A Monte Carlo method to estimate cell population heterogeneity

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

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

*ben.c.lambert@gmail.com.

Revision date & time: 2019-04-27 16:46

1 Abstract

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

2 Introduction

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

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

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

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

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

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

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

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

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

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

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

3 Method

In this section, we describe the first describe the probabilistic framework that underlies the CMC algorithm, before introducing CMC in pseudocode (Algorthm X). We also detail the workflow we have found useful in applying this approach to analyse cell snapshot data as well as suggest practical remedies to issues we have encountered in using CMC.

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

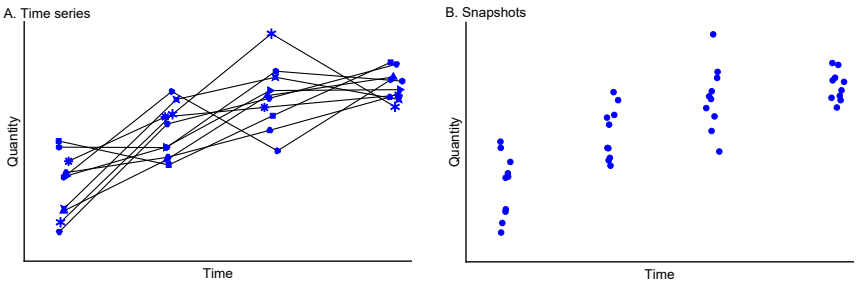

Figure 1: **Time series data (A.) versus snapshot data (B.) typical of single cell experiments.** In A. that the cell identities are retained at each measurement point (indicated by given plot markers) whereas in the snapshot data in B., this information is lost.

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

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

Here $\boldsymbol{X}(t) = (X_1(t), X_2(t), ..., X_k(t))$ is a vector of states for each compartment in the system and $f(.)$ is a function of these states and parameters $\boldsymbol{\theta} \in \mathbb{R}^p$. Note that in most circumstances, the initial state of the system, $\boldsymbol{X}(0)$, is unknown and it is convenient to include these as elements of $\boldsymbol{\theta}$ to be estimated.

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

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

Raw snapshot data consists of measurements of individual cells with exact inference requiring simulating the underlying ODE system for each individual. This is cumbersome and impractical for the numbers of cells sampled in typical experimental setups and so, instead, we follow previous work and instead represent snapshot data using probability distributions [12, 15–17]. The snapshots themselves can either be distributions of a single species or multiple species, which can be approximated by univariate and multivariate probability distributions respectively (Figure X).

Two panels: one shows independent snapshots at different points in time; the other, a multivariate snapshot.

3.1 Generative model

Contrast traditional Bayesian approach to ODE models (via figure already done).

3.2 Bayes' rule for HODEs

3.3 Contour Monte Carlo (CMC) and workflow

Figure showing workflow in using CMC to infer cell-to-cell variability.

4 Results

4.1 Growth factor model

Dixit simplified growth factor model. Target a unimodal distribution and show using both uniform and non-uniform priors.

4.2 Michaelis-Menten kinetics

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

4.3 TNF signalling pathway

Target a bimodal target.

5 Discussion

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

Natural selection has dictated that organisms have often evolved redundancies for systems essential for life. By building mathematical models we

aim to mimic these systems, meaning that these models should embody similar redundancies as their biological counterparts. Assessing the sensitivity of key characteristics of model outputs to perturbations in input parameters provides insight into the sensitivity of the system to each of its constituent elements. These so-called sensitivity analyses of mathematical models allow us to probe the biological system even when biological experiments are infeasible. Inverse sensitivity analysis inverts this process and instead of determining how model outputs vary in response to changes to the input parameter values, estimates a distribution over inputs which achieves a given distribution over outputs.

In this paper, we introduce an approach to inverse sensitivity analysis which can be applied to systems with many input parameters, mitigating the curse of dimensionality that limits the scope of methods which rely on grid-based approaches to build an explicit output-to-input map [18]. We have demonstrated that our algorithms can perform inverse sensitivity analysis on mathematical models of biological systems across a range of complexities, including the logistic growth model (2 inputs & 1 output target), Michaelis-Menten kinetics (3 inputs & 2 output targets), and the SIR model with uncertain initial population sizes (9 inputs & 3 output targets). As well as detailing our algorithms, we provide a probabilistic framework for understanding inverse sensitivity analysis, in which "prior" probability distributions are set on the inputs. These prior beliefs over input values are consistent with a "posterior" input distribution which, when transformed through the input-to-output map, results in a "target" output distribution. To sample from these posterior input distributions, we introduce a two-step sampling algorithm. In the first of these steps, input parameters are independently sampled from their prior distributions and, by fitting a kernel density estimator to the output values, this provides an approximate Jacobian transform (which we interchangeably term a "contour volume distribution"), which is used in the second step involving Markov chain Monte Carlo. A similar algorithm for inverse sensitivity analysis has recently been derived from a measure theoretic perspective by [19]. These authors also investigate stability of the posterior distribution with respect to the observed output distribution, the assumed prior distribution and the approximation of the contours of the forward map. We believe that the different path we take to the shared goal offers complementary insight into the algorithm's mechanism and provides an intuitive way to understand inverse sensitivity analysis, more generally.

There are several subtleties in the first steps of the processes described in Algorithms **??** and **??**) which must be understood in order to ensure a valid input distribution is obtained. Indeed, these intricacies complicated our own efforts in testing the algorithms. Provided the output is well-behaved over the space of possible input values, a univariate output distribution can be approximated given a relatively modest number of samples from the input priors using standard kernel density estimation (KDE). Here, for the univariate output target distributions we found that KDE with a Gaussian kernel using default bandwidths from each software package used (Matlab, Mathematica and R [20–22]) was able to represent the output distribution with sufficient fidelity to ensure the input posterior recaptured the output target. The number of input samples necessary to ensure convergence to the true posterior input distribution, however, depends on the exact output distribution being targeted. If the bulk of probability mass for the target output distribution lies at a location in output space where the

contour volume is rapidly varying, then the input distribution obtained will be sensitive to errors in kernel density estimates of the contour volume distribution, and many samples will be required. Similarly, if a region of low contour volume is targeted, then kernel density estimates with few samples will be relatively noisy and more samples will be necessary. Here we have assumed numerical errors in solving the map are negligible and independent of the parameters. Neither assumption is likely to be true for sophisticated partial differential equation models and the interaction between numerical and sampling errors is the subject of ongoing analysis. KDE introduces a further source of error, which must be carefully managed to ensure reasonable results are obtained.

Our algorithms avoid the curse of dimensionality of the input space which plagues grid-based approaches to inverse sensitivity analysis. The necessity of having to fit a probability distribution to the output samples resultant from sampling the prior input distributions means that, at present, our approach is limited to problems with relatively few outputs. In §**??**, the output target was a three-dimensional distribution, and we expended considerable effort finding a KDE method that adequately approximated the three-dimensional contour volume distribution. We ultimately found that the most effective approach was obtained by using the "kde" function within the "ks" R package [22, 23], which uses the data to estimate unconstrained bandwidth matrices, which are then used to fit kernel density estimates to data with up to six dimensions. Density estimation, however, is currently an active area of research and software packages exist implementing many different variants of KDE (see [24] for a review of the R packages already available in 2011). Vine copulas have recently been suggested as an approach which avoids the curse of dimensionality in density estimation [25]. If this promise is realised, then our algorithm will be applicable to output target distributions of higher dimensions.

Mathematical models have proved indispensable tools for elucidating understanding of biological systems, which are frequently not amenable to direct experimentation. Biological systems are often robust to perturbations to particular constituent processes, and we can use mathematical models to explore these sensitivities. Inverse sensitivity analyses are a relatively recent addition to a modeller's toolbox, which allows one to determine an input distribution - consistent with prior beliefs - that can generate a given distribution of outputs. Here we introduce a Monte Carlo method which extends the range of models for which inverse sensitivity analysis can be performed, and illustrate its utility for several problems of interest to computational biology. It is our hope that, by publishing this method, others are encouraged to undertake inverse sensitivity analysis, which we have found is insightful for building and analysing mathematical models.

6 Author contributions

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[18] T. Butler, D. Estep, S.J. Tavener, C. Dawson, and J.J. Westerink. A measure-theoretic computational method for inverse sensitivity problems iii: Multiple quantities of interest. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):174–202, 2014.

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

[20] Matlab version 9.0.0.341360 (r2016a). `https://www.mathworks.com/`, 2016.

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

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

[23] Tarn Duong and Maintainer Tarn Duong. Package ks. `https://cran.r-project.org/web/packages/ks/ks.pdf`, 2018.

[24] Henry Deng and Hadley Wickham. Density estimation in r. `https://vita.had.co.nz/papers/density-estimation.pdf`, 2011.

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