Response to referee – notes from Simon

**Response to editor** (cover letter)

COVID delay – two coauthors heavily involved with UK efforts

Difficulty in finding experimental data and suitable model. Most of the papers in the literature use simulated data. Nondimensionalization information somewhat incomplete. Valuable experience.

**Response to referee**

**Reviewer #1:**

MAJOR COMMENTS

(A) Please describe the HODE approach in more detail in the introductory section.

We have expanded the discussion in the introduction.

(B) The introduction would also benefit from an overview discussing different approaches of modelling cell populations with heterogeneous cell properties and from providing respective examples. E.g,

-Chan YH et al A subpopulation model to analyze heterogeneous cell differentiation dynamics. Bioinformatics. 2016;32(21):3306-3313

It is difficult to provide a comprehensive review of other methods for analysis and this is not the purpose of our article. We have referenced a number of competing approaches. We thank the referee for the Chan et al paper which led us to the data for our final example.

(C) The manuscript would clearly benefit from showing the application of the proposed framework to real data. I suggest to choose a dataset from public repositories and apply the algorithm to it.

This was indeed illuminating. Please see section 4.5, where we apply our technique to experimental data from a study of embryonic stem cell development.

(D) Instead of referring to a companion manuscript, please extend the reasoning in section 3.2. As it is now, this part is difficult to follow.

We have revised and strengthened this section, providing a more concise development than that appearing in the companion manuscript.

(E) The discussion would benefit from explicitly discussing potential drawbacks of the presented framework as well as difficulties that can be expected when applying it to real datasets.

This is now discussed in the new section 4.5

(F) The advantages of fitting densities to raw snapshot data and relying on the densities during the estimation process are obvious. Please also discuss potential drawbacks.

???

(G) Please explain the following statements in more detail and theoretical depth:

-"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 target output distribution [25]"

We have modified this statement and made clear this is true for any sampler.

-"Throughout the course of development of CMC, we have tested many KDE methods and have found vine copula KDE is best suited to approximating the higher dimensional probability distributions required in practice."

Other commonly available KDE methods perform poorly as dimension increases.

-"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"

???

- "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." this statement requires much more explanations.

We disagree. This is simply a question of having more parameters than measured quantities

-"Asymptotically (in terms of the sample size of both sampling steps), CMC produces a sample of parameter values (<theta> [1] , <theta> [2] , ...) which, when mapped to the output space, corresponds ..."

This is a property of MCMC algorithms

-Optional: A more detailed explanation of the theoretical framework could make the text more readable for biologists. The text as it is now is inaccessible to non-statisticians.

MINOR COMMENTS

- On page 4 the authors discuss existing approaches for inference based on HODE models. It would be niche to shortly mention advantages and disadvantages of each presented approach.

This is beyond the scope of the current paper and calls for a review manuscript

-"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 [27]." Please shortly state why.

???  
  
**Reviewer #2:**

MAJOR COMMENTS

(1) Your method for estimating cellular heterogeneity seems to be entirely based on the assumption that no measurement noise is present in the data. This is an assumption that can be made with most other methods for estimating variances and as such is not particular to this method. You should change your text and highlights to reflect that, or alternatively state how your method is otherwise more suited to estimating cellular variability than previously existing methods.

???

(2) You should explain the novelty of your method compared to existing Bayesian approaches better. At present time it is not clear to me how this method is better than the other methods you mention.

???  
  
(3) To showcase the differences to existing methods, it would be valuable for you to use existing methods on the problems in the Results section and compare the results to the results from your method.

Recoding existing methods is a complex issue, even when well described and is beyond the scope of this work. We agree with the referee, but this is a topic for a potential review article and also a major rationale for the PINTS project, with which two of the authors are associated which seeks to robustly implement a number of competing methods and make them available to users.

(4) While the applications shown on synthetic data make it clear that the method works on synthetic data, it would be interesting to see if this still holds up in real-world data. As such I suggest that you should apply your method to a real-world problem.

See new section 4.5

(6) The current explanation of iso-output contour regions (lines 254--255) is not clear enough. You should consider expanding on this concept to make it more clear.

We have expanded the discussion

(5) The explanation of the core part of this method is not clear. Specifically, how you arrived at equations 10 and 11 seems to be explained in a companion piece, rather than the text. To improve clarity in this section, you should give an expanded explanation of these equations and how you got here.

We have expanded the discussion

(6) How you found that "vine-copula KDE is best suited to approximating the higher dimensional probability distributions" (lines 334--336) is not mentioned. You should consider expanding on this.

We have expanded the discussion (see also referee #1).

(7) This text is not easy to follow for biologists. Addressing the above points is likely to make it easier to read. Even then, you should re-read your text with a biologist reader in mind.  
  
MINOR COMMENTS

(8) You should clarify why the property of MCMC to be biased towards regions of parameter space with larger iso-output contour volumes (lines 274-277) is not desired.

???  
  
(9) The language is in large parts inaccessible to biologists. You should consider introducing terms such as "Target distribution" before using them.

We have recognized this particular deficiency and added definitions.

(10) It was initially not clear that the fixed parameter set in the growth factor model was only fixed for data generation. You should clarify that it was not fixed in the subsequent steps.

???

(11) You should explain how the parameters for the Gaussian priors were chosen.

???  
  
(12) The use of this method is not clear for biologists.

(13) The julia "solve" method is actually not inbuilt, but part of the DifferentialEquations.jl library (line 381)