**Response to referees**

* Referee’s comments appear as blue text
* Our responses appear as black text
* Sentences or paragraphs incorporated in the manuscript appear as green text

**Reviewer #1**

MAJOR COMMENTS

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

We thank the reviewer for pointing this out: before we had not actually discussed the HODE system (nor explicitly defined) it in our methods. Now, we do so in Section 3.1:

“The collection of such idiosyncratic ODEs across all cells is then referred to as the ``HODE model''.”

(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

We thank the referee for the list of references: the Chan et al. paper, in particular, we found very relevant and have added the following to our introduction,

“The first involves using population average data -- mean values of measurements at different points in time -- yet, explicitly modeling how this mean represents a mixture across different subpopulations. 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 Chan paper was also useful as it led to the newly included real data example, we discuss below.

(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 4.4, 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.

The way in which we introduce CMC here mirrors what is done in the companion piece (and, since it postdates it, we think we’ve actually improved on our description there). What the companion paper does do, however, is provide many worked examples that further illustrate how the elements of the overall algorithm fit together. These worked examples are somewhat abstract, and we’d like to keep them separate from this “applications” paper. Such a discussion here would result in a much longer paper which we feel would detract from its readability and intended message.

(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 4.4. In particular, in fitting a literature model to these real data, we find it provides a relatively weak representation of the data generating process. The problem could, in theory, be due to the inability to include measurement noise in our current framework. (As such, extending our method to handle measurement noise would be a fruitful direction for future work.) We suggest, however, that, in this case, it is more likely this is due to ODE model misspecification given the extent of the discrepancy between forward model simulations and data.

“Indeed, this failure to target both q1 and q2 simultaneously suggests that the model does not actually cohere with the data… 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.”

(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.

This comment and the Chan reference caused us to restructure this part of our introduction. We have added the following paragraph that explicitly discusses the cost of fitting to densities, as opposed to raw 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 ~10^4 (Hasenauer et al., 2011), 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 modeling how this mean represents a mixture across different subpopulations. 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 (Hasenauer et al, 2011; Hasenauer et a., 2014; Loos et a., 2018; Dixit et al., 2018). We follow this approach here as it is likely that more information about the underlying data is retained than in the ``population average'' one.”

(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. We also emphasise again the difference between our problem and a traditional Bayesian one,

“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 (Lambert, 2018), 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 (Lambert et al., 2018). 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.”

-"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."

We have made explicit the reference for vine copula KDE, since this method was originally created with the specific purpose of dealing with high dimensional data:

“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 (Nagler et al., 2016).”

Nagler, Thomas, and Claudia Czado. "Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas." Journal of Multivariate Analysis 151 (2016): 69-89.

-"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"

This is a crucial point governing the existence of a solution to our inference problem. We have separated out this paragraph to illustrate this point. Rather than dwell further on abstract description of this aspect here, we have added to a cross-reference to a later point in the manuscript (Sections 4.2.2 and 4.4) where this is covered in the discussion of two of our numerical examples.

- "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.

The initial concentrations of species in cellular assays are measured quantities – that is, imperfect representations of the underlying quantities. As such, we prefer to estimate them through inference rather than fix them as we think this better reflects reality.

-"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 ..."

We have restructured this sentence, so that it’s easier to parse: “As the sample size of both sampling steps (i.e. the contour volume estimation and MCMC steps) tends to infinity…”

-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.

We believe that by addressing most of the comments of both referees, the text should be more readable for biologists. We recognise, however, that the topic is mathematically rich and, therefore, appropriate for the Journal of Theoretical Biology. We would suggest though that our exposition is, in general, less mathematically sophisticated than existing research papers on this area. At some point in the future, it may well be worth writing a review of these such methods aimed at specifically at biologists.

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.

We have restructured this aspect of our introduction somewhat (thanks to this reviewer’s comments covered above). Now, this section includes more discussion of the advantages and disadvantage of these various approaches.

-"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.

This is a good point and was a bit opaque in the previous manuscript. We have now added the following:

“…we use an adaptive MCMC algorithm to improve sampling efficiency [27].”

**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.

In the abstract, we explicitly state that our technique is for low noise systems, which we believe makes it clear from the onset the circumstances when this assumption is appropriate:

“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 isn’t to suggest this is a benefit of our approach, however. Indeed, we recognise that the assumption of no measurement noise is imperfect and, repeatedly throughout the manuscript, suggest that our method could be improved by allowing for measurement noise. Here’s an example of this in the discussion:

“Future work incorporating a stochastic noise process or, more generally, including stochastic cellular mechanisms is thus likely to be worthwhile.”

But, as we discuss below, the assumption of no measurement noise may, nonetheless, be reasonable enough for a given application:

“Our approach currently assumes that output variation is dominated by cellular variation in the parameter values of the underlying ODE, with measurement noise making a negligible contribution. Whether this is a reasonable assumption depends on the system under investigation and, more importantly, on experimental details. We recognise that neglecting measurement noise when it is, in fact, important in determining observed data means CMC will overstate cellular variation. It may also mean that some output distributions cannot be obtained by our model system (i.e. HODEs without noise).”

It's worth emphasising that there is no consensus regarding the correct way to perform inference for single cell data (which we discuss in our introduction): this is why such a variety of approaches exist and they all, ours included, have limitations. There are, however, a number of benefits of our approach to existing methods, which we believe merits its inclusion to the literature:

“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 Loos et al. 2018, CMC can fit multivariate snapshot data and unlike Dixit et al. 2018, does not use discrete bins to model continuous data.”

(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.

This is a fairly subtle distinction since a) our method is some senses Bayesian but b) the problem we face is sufficiently different to traditional Bayesian analyses that existing approaches cannot be used. Indeed, it took substantial thinking on our part to recast the problem in Bayesian terms.

We have attempted to address this issue by adding the following to our Methods, which we hope makes this distinction clearer:

“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 (Lambert, 2018), 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 (Lambert et al., 2018). 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.”

(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 these methods are unambiguously 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.

We have added a new section to address this point (see section 4.4). It was particularly challenging to obtain data here as none of the datasets that were previously used in the literature were both non-synthetic and freely available. Despite this, we eventually found an appropriate dataset to analyse, which we hope will be useful for others wanting to compare methodological approaches. Analysis of this dataset was illuminating and certainly strengthens the manuscript.  
  
(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.

These lines taken in isolation may be hard to parse but we think that Figure 3 (that explicitly gives an example of output contour volumes), mentioned in the following few sentences, explains this topic well.

(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.

The text currently details, we believe, the derivation of these equations in quite explicit and transparent form. The companion piece provides more detailed insight into the mathematical / geometrical form of these and, as a key point of difference from this work, provides somewhat abstract worked examples that illustrate how elements of the method function.

The current paper is focussed on the biology, so we’d prefer to avoid dwelling on these concepts here (and believe that we already have done so). Additionally, the companion piece is openly available, so anyone wanting to look into these further can do so.

(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 response to referee #1):

“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 (Nagler et al., 2016).”

(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.

See response to reviewer #1.

We believe that by addressing most of the comments of both referees, the text should be more readable to biologists. We recognise, however, that the topic is mathematically rich and, therefore, appropriate for the Journal of Theoretical Biology. We would suggest though that our exposition is, in general, less mathematically intense than existing research papers on this area. At some point in the future, it may well be worth writing a review of these such methods aimed at specifically at biologists.

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.

Any bias in the sampling distribution away from the posterior is undesirous since it would be a product of the sampling method *not* the inference problem. We explain this in the sentence which follows the one the reviewer highlights:

“This bias means that the stationary parameter distribution obtained, when fed through the model, does not recapitulate the target output distribution (Lambert et al., 2018).”  
  
(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 expanded our discussion of this term as follows:

“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, q in\R^m. We suppose these kernel density estimates approximate a true distribution over the observed data, p(q|Phi) and denote the estimated density as p(q|Phi\_hat). After this initial fitting, this distribution -- which we term the ``target distribution'' -- becomes the object we seek to replicate in our inference problem.”

(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.

We have now added the following to this part of the manuscript:

“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 S4.1.1 and S4.1.2 (where we use CMC to perform inference).”

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

The Gaussian priors were chosen to be “more concentrated than the uniform priors used in S4.1.1”. But in a sense are arbitrary. The example is intended to show that, since the inference problem is unidentified, arbitrary changes in the prior can affect the posterior. As such, we think it is fine to leave this section as it stands.   
  
(12) The use of this method is not clear for biologists.

See response to reviewer #1.

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

Thanks for pointing this out. We have added a reference to the paper where this package was introduced.