Simulation-based Evaluation of the Reliability of Bayesian Hierarchical Models for sc-RNAseq Data

Sijia Li
School of Mathematics
University of Leeds
Leeds, United Kingdom
sijia.li.ldn@gmail.com

Martín López-García
School of Mathematics
University of Leeds
Leeds, United Kingdom
M.Lopezgarcia@leeds.ac.uk

Luisa Cutillo
School of Mathematics
University of Leeds
Leeds, United Kingdom
L.Cutillo@leeds.ac.uk

***Abstract*—A Bayesian hierarchical model (BHM) is typically formulated specifying the data model, the parameters model and the prior distributions. The posterior inference of a BHM depends both on the model specification and on the computation algorithm used. The most straightforward way to test the reliability of a BHM inference is to compare the posterior distributions with the ground truth value of the model parameters, when available. However, when dealing with experimental data, the true value of the underlying parameters is typically unknown. In these situations, numerical experiments based on synthetic datasets generated from the model itself offer a natural approach to check model performance and posterior estimates. Surprisingly, validation of BHMs with high-dimensional parameter spaces and non-Gaussian distributions, is unexplored. In this paper, we show how to test the reliability of a BHM. We introduce a change in the model assumptions to allow for prior contamination, and develop a simulation-based evaluation framework to assess the reliability of the inference of a given BHM. We illustrate our approach on a specific BHM used for the analysis of Single-cell Sequencing Data (BASiCS).**

***Index Terms*—Bayesian Hierarchical Model; Single-cell Sequencing Data; Parameter calibration; Simulation-based Calibration**

I. Introduction

Bayesian Hierarchical Models (BHM) take into account relations between variables [1] by assuming a joint probability distribution for a set of parameters to be related to the observation of interest. Lately, BHMs have been used in biomedical applications and validating the reliability of BHMs with high-dimensional parameters is a challenging task, especially when applied to noisy biological data. *Single-cell RNA sequencing* (scRNAseq) is a recent technique to quantify RNA molecules on single-cell level, thus providing insights to the gene expression profile of each cell [2].

A recent example of BHM applied to scRNAseq data is the Bayesian Analysis of Single-Cell Sequencing Data (BASiCS) framework introduced in [3]–[5]. BASiCS aims to provide a structural method to analyse scRNAseq count data while separating various latent variables affecting gene expression, and therefore detecting gene expression heterogeneity in downstream analysis. In its early release, BASiCS was introduced as a non-regression model [3], [4], assuming independence between the mean and variance factors in the model. The latest version of BASiCS is presented as a regression model in [5], considering the confounding effect between mean and variance [6]. The downstream analysis in this framework depends on the posterior inference of the variables in the BHM.

In this paper, we examine the reliability of both the non-regression BASiCS [4] and the regression BASiCS [5]. Both BASiCS models propose the posterior median to estimate relevant variables in the downstream analysis. To validate this estimation, we work on synthetic datasets generated from the corresponding prior model. To explore the influence of prior specification, we also modify the original package to introduce a continuous range of choices for the prior distribution of the biological variation variable, testing the model robustness under perturbed prior models. Finally, we show how the Simulation-based Calibration method recently developed in [7] can be adapted here to validate high-dimensional parameter inferences.

II. BASiCS framework

BASiCS [3]–[5] aims to provide a structural method to analyse scRNAseq count data to detect gene expression heterogeneity. For a $q \times n$ scRNAseq count matrix, with n cells, q_0 biological genes and $q-q_0$ spike-in genes, BASiCS assumes the following likelihood model:

$$X_{ij} \,|\, \mu_i, \delta_i, \Phi_j, \nu_j \overset{ind}{\sim} \begin{cases} \text{Neg-Binomial}\left(\delta_i^{-1}, \frac{\Phi_j \nu_j \mu_i}{\Phi_j \nu_j \mu_i + \delta_i^{-1}}\right), \\ \qquad\qquad \textit{for } i \in \{1, \ldots, q_0\}, \\ \text{Poisson}\left(\nu_j \mu_i\right), \\ \qquad\qquad \textit{for } i \in \{q_0+1, \ldots, q\}, \end{cases} \tag{1}$$

where μ_i is the expected expression count for gene i across all cells, δ_i the biological expression heterogeneity variable for gene i, Φ_j the cell size variable of cell j, and ν_j the technical noise variable for cell j. The priors of μ_i, Φ_j and ν_j are $\mu_i \,|\, \sigma_\mu \overset{ind}{\sim} \text{log-Normal}\left(0, \sigma_\mu^2\right)$, $(\Phi_1, \ldots, \Phi_n) \,|\, \boldsymbol{p} \sim n\text{Dirichlet}(\boldsymbol{p})$ and $\nu_j \,|\, \theta, s_j \overset{ind}{\sim} \text{Gamma}\left(\frac{1}{\theta}, \frac{1}{s_j \theta}\right)$, with prior on hyperparameters, θ the global technical noise variable and s_j the cell-j-specific technical noise variable:

$$\theta \,|\, a_\theta, b_\theta \sim \text{Gamma}(a_\theta, b_\theta), \quad s_j \,|\, a_s, b_s \overset{i.i.d.}{\sim} \text{Gamma}(a_s, b_s).$$

This research was supported by an EPSRC Doctoral Training Partnership (reference EP/R513258/1) through the University of Leeds.

What differentiates the Non-regression BASiCS [3], [4] model and the Regression BASiCS model [5] is the prior assigned to the biological expression heterogeneity variable δ_i.

a) Non-regression BASiCS model:

$$\textbf{Log-normal prior:} \quad \delta_i \,|\, \sigma_\delta \overset{ind}{\sim} \text{log-Normal}\left(0, \sigma_\delta^2\right), \quad (2)$$

$$\textbf{Gamma prior:} \quad \delta_i \,|\, a_\delta, b_\delta \overset{ind}{\sim} \text{Gamma}\left(a_\delta, b_\delta\right), \quad (3)$$

b) Regression BASiCS model:

$$\delta_i \left| \mu_i, \beta, \sigma_\delta^2, \lambda_i \right. \overset{ind}{\sim} \text{log-Normal}\left(f(\mu_i), \frac{\sigma_\delta^2}{\lambda_i} \right), \qquad (4)$$

where $f(\mu_i)$ is a nonlinear regression of μ_i. For more details on BASiCS framework, see the Appendix.

III. Simulation-based Evaluation of BHMs

For BHMs applied to biological data, it is rare to have the ground truth of latent variable values to assess the recovery of parameters of interest. In this study we simulate the gene expression count matrix from the prior model of non-regression and regression BASiCS respectively, and then we use the synthetic data in the corresponding BASiCS MCMC to compare the estimated posteriors with the "true" parameter values used for data generation. For a detailed description of the prior models, please see the Appendix. Our experiments are performed in `R 4.0.2` [8], code available at https://github.com/lilythepooh/BASiCS-Reliability.git.

A. Uncertainty of median as a point estimate

In [3]–[5], the posterior medians are proposed as estimates for δ_i, μ_i, ν_j, Φ_j, s_j and θ, for downstream analysis. We simulate a dataset $\mathbf{X}^{(1)*}$ of 100 genes, 10 spike-in genes, and 50 cells from non-regression BASiCS model [4], simulating δ_i from log-Normal distribution as in Eq. (2). After the required data-preprocessing procedure of BASiCS, the resulted dataset of synthetic gene expression has $n = 39$ cells, $q_0 = 100$ biological genes and $q - q_0 = 10$ spike-in genes. Then we plug this synthetic dataset into the Monte Carlo Markov Chain (MCMC) algorithm for non-regression BASiCS (`BASiCS` package [4]), with fixed prior-hyperparameter values.

Similarly, we simulate a dataset $\mathbf{X}^{(2)*}$ of 100 biological genes, 10 spike-in genes and 50 cells from regression BASiCS model [5], where δ_i was simulated from log-Normal distribution as in Eq. (4). After the data preprocessing procedure required for BASiCS, we get a synthetic gene expression count dataset with $n = 44$ cells, $q_0 = 98$ biological genes and $q - q_0 = 10$ spike-in genes. Then we plug the fixed synthetic dataset back into the MCMC algorithm for regression BASiCS within the `BASiCS` package [5], with fixed prior-hyperparameter values. When recovering the parameter values used for generating datasets $\mathbf{X}^{(\mathbf{1})*}$ and $\mathbf{X}^{(\mathbf{2})*}$, we replicate 100 MCMC respectively, resulting in 100 posterior medians for each parameters μ_i and δ_i from each model, $i = 1, \ldots, q_0$. Each of the 100 MCMC was run for 15,000 iterations, 10,000 burns and thinned by 5, resulting in 100 posterior samples of size 1,000 for both models respectively.

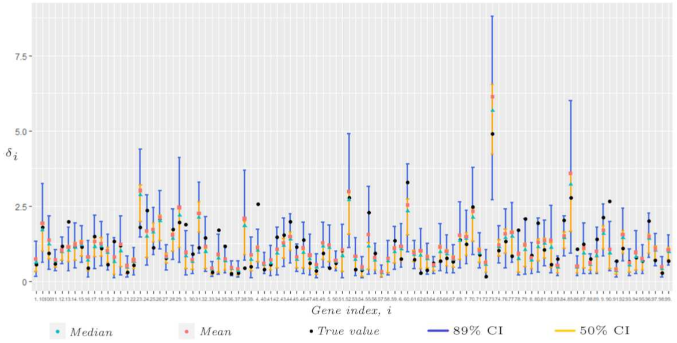

Fig. 1: True values (δ_i^*) and posterior estimates (median, mean, 89% and 50% Highest Density Credible Intervals) for δ_i, for biological genes $i = 1, \ldots, q_0$ ($q_0 = 100$) and for one particular replication of the estimation procedure. Non-regression BASiCS model.

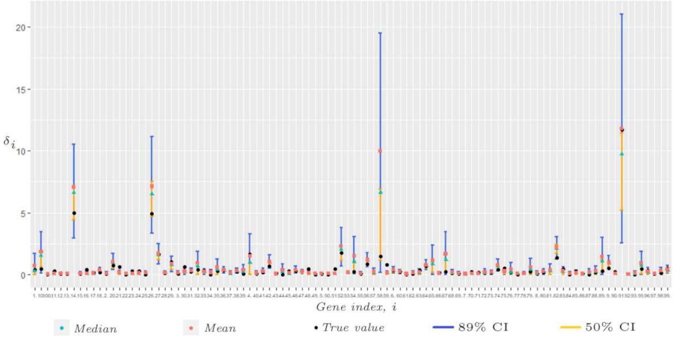

Fig. 2: True values (δ_i^*) and posterior estimates (median, mean, 89% and 50% Highest Density Credible Intervals) for δ_i, for biological genes $i = 1, \ldots, q_0$ ($q_0 = 98$), and for one particular replication of the estimation procedure. Regression BASiCS model.

To illustrate the recovery of true parameters in each run, we calculated the 89% Highest Density Credible Interval using `bayestestR` package [9]. We calculated the 89% Credible Interval rather than the more common 90% or 95% because according to [9], 89% credible intervals are typically more stable. Figure 1 shows that from non-regression BASiCS, among 100 gene-specific biological variation parameters $\{\delta_i : i \in \{1, \ldots, 100\}\}$, 11 of the true values do not fall inside the 89% Highest Density Credible Interval, and 53 of the true values do not fall inside the 50% Highest Density Credible Interval. As can be observed, the level of stochasticity in this BHM means that using the posterior median as a single estimate of the parameter does not necessarily work well, since these single estimates may not even properly capture the relative relationship between two δ_i's for two different genes. For example, δ_{80} and δ_{82} have very similar posterior median and posterior mean estimates (around 1.25), but the true value for gene 82 ($\delta_{82}^* = 0.561$) is much smaller than for gene 80 ($\delta_{80}^* = 1.944$), indicating a lower biological variation factor for gene 82. We note here that our comments above remain after exploring other 199 runs of the estimation process, since they are replication runs of the same model using the same dataset under the same conditions.

Figure 2 shows that from regression BASiCS, among 98 gene-specific biological variation parameters δ_i, 10 of the true values do not fall inside the estimated 89% Highest Density Credible Interval, and 48 of the true values do not fall inside the 50% Highest Density Credible Interval. We note that those 10 true values outside of the 89% Highest Density Credible Intervals are small values between $(0,1)$, leading to very narrow posterior Highest Density Credible Intervals. In this case, the posterior median could still act as a fair single point estimate for them. For most δ_i, the variance of the posteriors looks much smaller compared to Figure 1, but such precision only occurred on the posteriors of δ_i with small true values. When simulated from regression BASiCS model, most biological variation factors δ_i are small.

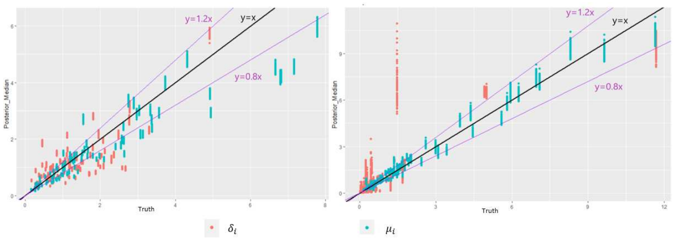

Fig. 3: True value VS posterior median of all gene-specific parameters δ_i and μ_i, with 100 replications and fixed dataset. **Left.** Results from non-regression BASiCS with log-normal prior for δ_i. **Right.** Results from regression BASiCS.

Figure 3 shows that the posterior medians vary across different MCMC runs, and some of them are not in the 20% relative error range of the true values. In regression BASiCS, the posterior median of μ_i is closer to the truth, but the posterior median of δ_i varies more across different MCMC runs.

B. Posterior Predictive Check

Following [10], we use Posterior Predictive Check (PPC) to assess our model fit. For a run in Section A, from each set of parameters $\{\mu_i, \nu_j, \Phi_j, \delta_i\}$ in 1000 posterior samples, we simulate a posterior predictive value of biological gene expression count $X_{ij}^{(1)}$ ($X_{ij}^{(2)}$) from Eq. (1), resulted in 1000 posterior predictive $X_{ij}^{(1)}$ ($X_{ij}^{(2)}$) to compare with the true data $X_{ij}^{(1)*}$ ($X_{ij}^{(2)*}$). Figure 4 and Figure 5 plot the histogram of posterior predictive $X_{ij}^{(1)}$ ($X_{ij}^{(2)}$) and the vertical line of $x = X_{ij}^{(1)*}$ ($x = X_{ij}^{(2)*}$) for non-regression BASiCS model and regression BASiCS model respectively. We can see that the regression BASiCS model [5] has performed better compared with the non-regression BASiCS model. [4].

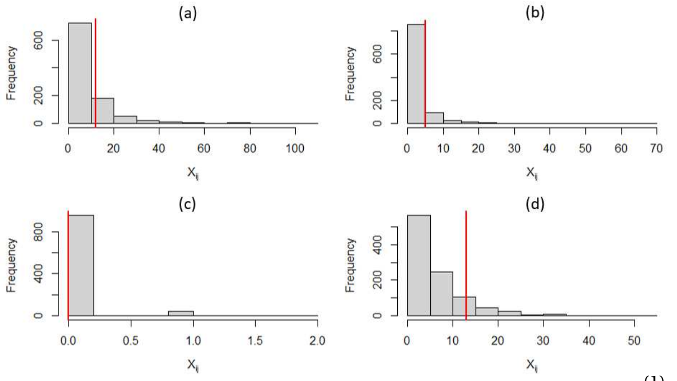

Fig. 4: **histogram:** posterior predictive distribution of $X_{ij}^{(1)}$, simulated from the posteriors of run 1.
red line: input data $X_{ij}^{(1)*}$.
(a): gene $i = 7$, cell $j = 1$. **(b):** gene $i = 10$, cell $j = 12$.
(c): gene $i = 38$, cell $j = 14$. **(d):** gene $i = 56$, cell $j = 33$.

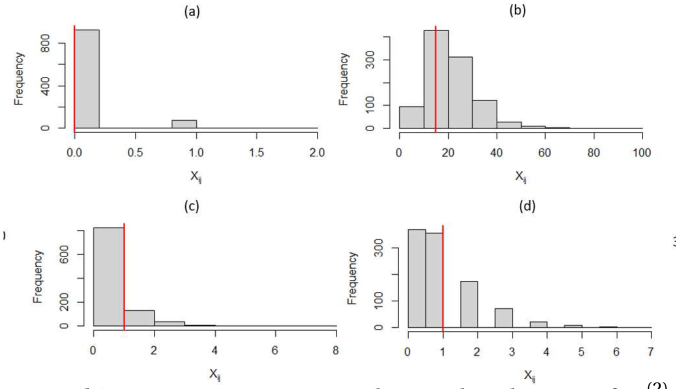

Fig. 5: **histogram:** posterior predictive distribution of $X_{ij}^{(2)}$, simulated from the posteriors of run 1, regression BASiCS model.
red line: input data $X_{ij}^{(2)*}$.
(a): gene $i = 1$, cell $j = 9$. **(b):** gene $i = 12$, cell $j = 2$.
(c): gene $i = 15$, cell $j = 11$. **(d):** gene $i = 77$, cell $j = 7$.

C. Sensitivity to contamination on prior

To assess the sensitivity of the Bayesian Hierarchical model to prior choices, we introduce a contamination on prior distribution. In [3], the prior distribution of the gene-specific biological variation variable δ_i is a Gamma distribution, while in [4] the prior distribution of δ_i is a log-Normal distribution. Given the expected limited information on δ_i, we propose the following extension of the prior specification of δ_i in non-regression BASiCS model:

$$\delta_i \overset{ind}{\sim} (1-\varepsilon)\cdot\text{log-Normal}\left(0,\sigma_\delta^2\right)+\varepsilon\cdot\text{Gamma}\left(a_\delta, b_\delta\right), \quad \varepsilon \in [0,1], \tag{5}$$

where ε is the proportion of the Gamma prior in the prior mixture. This extends the bipolar choice in the original package to a continuous range of choice for the prior distribution of δ_i. We modified accordingly the non-regression part of the BASiCS package [4], to explore how different choices of the prior family affect the posterior inference. We simulated a synthetic dataset $\mathbf{X}^{(1)*}$ with $\varepsilon = 0$ in Eq (5), with everything else follows the original non-regression BASiCS model. Using the synthetic dataset $\mathbf{X}^{(1)*}$, we apply our modified MCMC with fixed prior-hyperparameter values and $\varepsilon \in \{0, 0.5, 1\}$. Here $\varepsilon \in \{0, 0.5, 1\}$ is an example to explore how different mixtures of prior model could affect the posterior result. To investigate the stochastic variation in the MCMC result, for each $\varepsilon \in \{0, 0.5, 1\}$ we replicate the MCMC for 200 times.

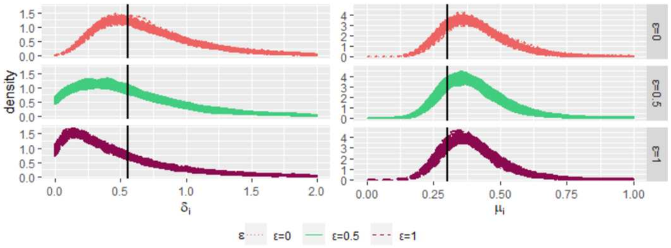

Fig. 6: Gene-specific estimated posterior distributions for δ_1 and μ_1 from the synthetic dataset. For δ_1, a "true" value $\varepsilon^* = 0$ was considered to generate the dataset. The "true" values of δ_1^* and μ_1^* are represented by vertical lines. During the estimation process, we consider different prior distributions for δ_1 by varying $\varepsilon \in \{0, 0.5, 1\}$. 200 replications are considered for each ε choice.

In Figure 6, we plot the posterior samples for the parameters δ_1 and μ_1, for any $\varepsilon \in \{0, 0.5, 1\}$. It is easy to see that the true parameters δ_1^* and μ_1^* are successfully recovered within the posterior ranges. Still, a certain degree of stochasticity from the MCMC estimation leads to variability in the posterior samples around those true values. On the other hand, it seems clear that the estimation of δ_1 worsens for increasing values of ε. This is expected, since the true value of ε used to simulate the synthetic dataset $\mathbf{X}^{(1)*}$ was $\varepsilon^* = 0$. These results suggest that the recovery of μ_1 seems to remain largely unaffected by changes in ε, which is to be expected since μ_1 is independent from ε. On the other hand, the selection of the prior distribution for δ_i (either Gamma, log-Normal or something in between) can have a noticeable impact on the corresponding estimates. However, this is not always the case with δ_i.

Table I shows the posterior median and the posterior credible intervals for δ_i and μ_i corresponding to several different genes $i \in \{1, 38, 39, 60\}$ in non-regression BASiCS. These genes have been chosen accordingly here to illustrate different behaviours. In Table I, the posteriors of μ_i are consistent regardless of the choice of ε. While the posterior medians of μ_1, μ_{38} and μ_{39} are relatively close to the true values, and the confidence intervals typically contain the true value inside, the true value μ_{60}^* does not fall into the posterior 89% Highest Density Credible Interval, and μ_{60} is underestimated in this case.

ε	δ_1		δ_{38}		δ_{39}		δ_{60}	
	Median	89% CI	Median	89% CI	Median	89% CI	Median	89% CI
0	0.64	[0.16,1.32]	1.84	[0.71,3.70]	0.73	[0.11,1.46]	2.33	[0.97,3.90]
0.5	0.48	[0.00,1.14]	1.87	[0.63,3.41]	0.56	[0.00,1.42]	2.30	[0.87,3.88]
1	0.35	[0.01,0.96]	1.90	[0.73,3.66]	0.40	[0.00,1.39]	2.40	[0.92,3.96]
Truth	$\delta_1^* = 0.56$		$\delta_{38}^* = 0.45$		$\delta_{39}^* = 0.50$		$\delta_{60}^* = 3.29$	

ε	μ_1		μ_{38}		μ_{39}		μ_{60}	
	Median	89% CI	Median	89% CI	Median	89% CI	Median	89% CI
0	0.39	[0.20,0.56]	0.43	[0.20,0.68]	0.21	[0.11,0.36]	0.69	[0.32,1.08]
0.5	0.38	[0.22,0.55]	0.42	[0.18,0.66]	0.20	[0.09,0.32]	0.67	[0.33,1.07]
1	0.38	[0.23,0.56]	0.43	[0.19,0.72]	0.21	[0.11,0.32]	0.69	[0.33,1.13]
Truth	$\mu_1^* = 0.30$		$\mu_{38}^* = 0.32$		$\mu_{39}^* = 0.18$		$\mu_{60} * = 1.30$	

TABLE I: Posterior medians and 89% Highest Density Credible Intervals of δ_i and μ_i, $i \in \{1, 38, 39, 60\}$, for different choices of $\varepsilon \in \{0, 0.5, 1\}$. The "true" values μ_i^* and δ_i^* for each gene used to generate the dataset are also reported.

In Table I it can be observed how the posteriors of δ_1 and δ_{39} are more sensitive to the choice of ε, while the posteriors of δ_{38} and δ_{60} are more robust against this choice. On the other hand, although the true values δ_1^*, δ_{39}^* and δ_{60}^* fall inside the posterior 89% Highest Density Credible Interval, the posterior median as a single estimate is relatively far away from the true value. Even when considering the log-Normal prior ($\varepsilon = 0$, the value used for generating the dataset), δ_{39} is still overestimated, and δ_{60} is underestimated. Moreover, in cases like δ_{38}, the MCMC recovery of δ_{38} is consistently wrong, regardless of the choice of ε, i.e. the prior distribution. In particular, not only the posterior median is far from the true value δ_{38}^*, but also δ_{38}^* does not fall inside the posterior 89% Highest Density Credible Interval. To understand the reason behind this, we look at the simulated gene expression count of gene 1 and gene 38. In the dataset $\mathbf{X}^{(\mathbf{1})*}$, the gene expression count of gene 1 $X_{1j}^{(1)*}$ has 29 zero values out of 39. The 10 non-zero $X_{1j}^{(1)*}$ range in $\{1, 2, 3, 4\}$. While in dataset $\mathbf{X}^{(\mathbf{1})*}$, the gene expression counts of gene 38 across cells, $\{X_{38,j}^{(1)*} : j = 1, \ldots, 39\}$, contain 32 zero values out of 39. The 7 non-zero values range within the set $\{1, 2, 3, 6\}$. Therefore, in synthetic dataset $\mathbf{X}^{(\mathbf{1})*}$, there is more variation across cells for the gene expression count of gene 38 $X_{38,j}^{(1)*}$ across cells compared to gene 1. Thus, the estimation process would naturally conclude that gene 38 has a higher value of δ_{38}, hence the consistent overestimation of δ_{38}.

D. Simulation-Based Calibration Adapted for BHM with High-dimensional Parameters

Simulation-based Calibration (SBC) is a general procedure proposed in [7] for validating inferences from Bayesian algorithms capable of generating posterior samples. Consider a joint distribution over measurements $\boldsymbol{x}$ and parameters $\boldsymbol{\zeta}$, with specified likelihood $\pi(\boldsymbol{x}|\boldsymbol{\zeta})$ and prior distribution $\pi(\boldsymbol{\zeta})$, so that $\pi(\boldsymbol{x}, \boldsymbol{\zeta}) = \pi(\boldsymbol{x}|\boldsymbol{\zeta}) \cdot \pi(\boldsymbol{\zeta})$. Bayes' Theorem yields that for a set of observations $\tilde{\boldsymbol{x}}$, the posterior distribution $\pi(\boldsymbol{\zeta}|\tilde{\boldsymbol{x}}) \propto \pi(\tilde{\boldsymbol{x}}, \boldsymbol{\zeta})$. Denote the corresponding parameter space of $\boldsymbol{\zeta}$ as $\mathscr{Z}$. Suppose we simulate a ground truth $\tilde{\boldsymbol{\zeta}} \in \mathscr{Z}$ from the prior $\tilde{\boldsymbol{\zeta}} \sim \pi(\boldsymbol{\zeta})$, and then generate some data from the corresponding data generating process $\tilde{\boldsymbol{x}} \sim \pi\left(\boldsymbol{x}|\tilde{\boldsymbol{\zeta}}\right)$. It is clear that, by integrating the exact posteriors over the Bayesian joint distribution, one gets the prior distribution

$$\pi(\boldsymbol{\zeta}) = \int \pi\left(\boldsymbol{\zeta}|\tilde{\boldsymbol{x}}\right) \pi\left(\tilde{\boldsymbol{x}}|\tilde{\boldsymbol{\zeta}}\right) \pi\left(\tilde{\boldsymbol{\zeta}}\right) \mathrm{d}\tilde{\boldsymbol{x}} \mathrm{d}\tilde{\boldsymbol{\zeta}}. \tag{6}$$

Eq. (6) is called the self-consistency condition in [7]. Consider drawing a sequence of samples from the posterior distribution $\pi(\boldsymbol{\zeta}|\tilde{\boldsymbol{x}}) \propto \pi(\tilde{\boldsymbol{x}}|\boldsymbol{\zeta})\pi(\boldsymbol{\zeta})$: $\{\boldsymbol{\zeta}(1), \ldots, \boldsymbol{\zeta}(L)\} \sim \pi(\boldsymbol{\zeta}|\tilde{\boldsymbol{x}})$. Condition (6) implies that $\tilde{\boldsymbol{\zeta}}$ and $\{\boldsymbol{\zeta}(1), \ldots, \boldsymbol{\zeta}(L)\}$ will be distributed according to the same distribution. Based on this, [7] defined the rank statistic

$$r\left(g\left(\boldsymbol{\zeta}(1)\right), \ldots, g\left(\boldsymbol{\zeta}(L)\right), g\left(\tilde{\boldsymbol{\zeta}}\right)\right) = \sum_{l=1}^{L} \mathbb{1}_{\{\boldsymbol{\zeta}(l) \ \ s.t. \ \ g(\boldsymbol{\zeta}(l)) < g(\tilde{\boldsymbol{\zeta}})\}}(\boldsymbol{\zeta}(l)), \tag{7}$$

which can be defined for any one-dimensional random variable $g : \mathscr{Z} \mapsto \mathbb{R}$, and where for a set A, $\mathbb{1}_A(a) = 1$ if $a \in A$, and is equal to 0 otherwise. [7] showed that given an i.i.d.

sample $\{\boldsymbol{\zeta}(1),\dots,\boldsymbol{\zeta}(L)\}$ from the posterior and a function $g: \mathcal{Z} \mapsto \mathbb{R}$, the rank statistic in Eq. (7) follows a discrete uniform distribution on $\{0,1,\dots,L\}$. Based on the uniformity of the rank statistic, they introduced Simulation-based Calibration as a way of exploiting this result to validate the inference process in practice, by checking that the resulting rank statistic is uniformly distributed (see Algorithm 1 in [7]).

To eliminate the potential bias from the autocorrelation structure in the Monte Carlo Markov Chain (MCMC), [7] propose that for MCMC methods, one can add a step for assessing Effective Sample Size $N_{eff}[g]$ with respect to the measurement of interest $g(\boldsymbol{\zeta})$. If $N_{eff}[g] > L$ then the autocorrelation is negligible, otherwise the MCMC needs to be rerun for an appropriate length of iterations (see Algorithm 2 in [7]).

BASiCS is implemented via MCMC , therefore it can be assessed with the Effective Sample Size assessment. Notably, in complex real data models like BASiCS, we have multiple measurements of interest $g_1(\boldsymbol{\zeta}),\dots,g_M(\boldsymbol{\zeta})$. Therefore, similar to [7], we assess the minimal Effective Sample Size with respect to all the measurements of interest, that is, if:

$$\min_{m=1,\dots,M} \{N_{eff}[g_m]\} > L \tag{8}$$

To implement this approach for the BHM in Section II, we define g as the projection function to each individual parameter in the parameter vector $\boldsymbol{\zeta}$. This is similar to the identity function g proposed in [11] for models with a single parameter, where the diagnosis consists of checking if the rank statistic for the parameter mirrors a uniform distribution. However, in BHMs like BASiCS, the approach can be adapted for the high-dimensional parameter space [7] $\boldsymbol{\zeta} = (\delta_1,\dots,\delta_{q_0},\mu_1,\dots,\mu_{q_0},\nu_1,\dots,\nu_n,\phi_1,\dots,\phi_n,s_1,\dots,s_n,\theta)$. We define $g_{\delta_i}(\boldsymbol{\zeta}) = \delta_i$, $g_{\mu_i}(\boldsymbol{\zeta}) = \mu_i$, $g_{\nu_j}(\boldsymbol{\zeta}) = \nu_j$, $g_{\Phi_j}(\boldsymbol{\zeta}) = \Phi_j$,$g_{s_j}(\boldsymbol{\zeta}) = s_j$,$g_\theta(\boldsymbol{\zeta}) = \theta$ in Eq (6), and we assess if:

$$\min_{i,j} \{N_{eff}[\delta_i], N_{eff}[\mu_i], N_{eff}[\nu_j], \\ N_{eff}[\Phi_j], N_{eff}[s_j], N_{eff}[\theta]\} > L.$$

This leads to Algorithm 1, which we implement and apply to the BASiCS non-regression model. In order to check the deviation of rank statistics from Uniform($\{0,1,\dots,L\}$), we plot the empirical cumulative density function (ECDF) and the expected CDF behaviour of Uniform($\{0,1,\dots,L\}$).

As Algorithm 1 demonstrated, in $k=1,\dots,K$ runs, all the parameters $\widetilde{\delta_i^{(k)}}$, $\widetilde{\mu_i^{(k)}}$, $\widetilde{\nu_j^{(k)}}$, $\widetilde{\Phi_j^{(k)}}$,$\widetilde{s_j^{(k)}}$, $\widetilde{\theta(k)}$ are re-simulated from the corresponding *i.i.d.* prior distribution for all $i = 1,\dots,q_0$, $j = 1,\dots,n$. Therefore, the rank statistic of each δ_i is equivalent to each other, the same applies to $\mu_i,\nu_j,\Phi_j,s_j,\theta$. Without losing generality, in Figure 7, we plot the ECDF of $\delta_1, s_1, \theta, \Phi_1, \nu_1, \mu_1$. From Figure 7, one can observe that the behaviour of the rank statistics for most of the parameters are close to the uniform distribution. On the other hand, the rank statistics for θ are far from the uniform distribution, suggesting that θ is likely to be underestimated in this model.

Algorithm 1: SBC for BASiCS: individual parameters

```
Require: Data generating model $\pi(X_{ij}|\delta_i,\mu_i,\nu_j,\Phi_j,s_j,\theta)$,
  prior distribution $\pi(\delta_i),\pi(\mu_i),\pi(\nu_j),\pi(\Phi_j),\pi(s_j),\pi(\theta)$, the
  number of rank statistic $K$, , the number of MCMC
  iterations $L'$, the resulted posterior MCMC chain length
  $N_{sample}$, the number of posterior sample used for
  calculating each rank statistic $L \approx \frac{N_{sample}}{10}$.
Initialise
while $k$ in $(1:K)$ do
    Draw prior sample for i=1,...,q,;
    j=1,...,n:
    $\widetilde{\delta_i^{(k)}} \sim \pi(\delta_i)$, $\widetilde{\mu_i^{(k)}} \sim \pi(\mu_i)$, $\widetilde{\nu_j^{(k)}} \sim \pi(\nu_j)$, $\widetilde{\Phi_j^{(k)}} \sim \pi(\Phi_j)$,
    $\widetilde{s_j^{(k)}} \sim \pi(s_j)$, $\widetilde{\theta^{(k)}} \sim \pi(\theta)$;
    Draw a simulated dataset, for $i=1,\dots,q$, $j=1,\dots,n$:
    $\widetilde{X_{ij}^{(k)}} \sim \pi\left(X_{ij}|\widetilde{\delta_i^{(k)}},\widetilde{\mu_i^{(k)}},\widetilde{\nu_j^{(k)}},\widetilde{\Phi_j^{(k)}},\widetilde{s_j^{(k)}},\widetilde{\theta^{(k)}}\right)$;
    Run the corresponding MCMC algorithm with Input
    dataset $\widetilde{\mathbf{X^{(k)}}} = \left(\widetilde{X_{ij}^{(k)}}\right)$ in BASiCS package for $L'$
    iterations to generate the correlated posterior sample
    chain of length $N_{sample}$ from
    $\pi(\delta_i^{(k)},\mu_i^{(k)},\nu_j^{(k)},\Phi_j^{(k)},s_j^{(k)},\theta^{(k)}|y^{\tilde{(k)}})$:
    $\left(\delta_i^{(k)}(t),\mu_i^{(k)}(t),\nu_j^{(k)}(t),\Phi_j^{(k)}(t),s_j^{(k)}(t),\theta^{(k)}(t)\right)$ for
    $t=1,\dots,N_{sample}$, , $i=1,\dots,q$, $j=1,\dots,n$;
    Call R function LaplacesDemon::ESS [12] to
    compute the effective sample size for each parameter,
    $N_{eff}^{(k)}[\delta_i]$, $N_{eff}^{(k)}[\mu_i]$, $N_{eff}^{(k)}[\nu_j]$, $N_{eff}^{(k)}[\Phi_j]$, $N_{eff}^{(k)}[s_j]$,
    $N_{eff}^{(k)}[\theta]$, for $i=1,\dots,q$, , $j=1,\dots,n$.
      $N_{eff}^{(k)} = \min\{N_{eff}^{(k)}[\delta_i],N_{eff}^{(k)}[\mu_i],N_{eff}^{(k)}[\nu_j],N_{eff}^{(k)}[\Phi_j],$
            $N_{eff}^{(k)}[s_j],N_{eff}^{(k)}[\theta]\}$          (9)
    if $N_{eff}^{(k)} < L$ then
        rerun the MCMC for $\frac{L'\cdot L}{N_{eff}^{(k)}}$ iterations.
    else
        For each $i=1,\dots,q_0$, $j=1,\dots,n$, thin the posterior
        MCMC chain to L samples
        $\left\{\left(\delta_i^{(k)}(t_l),\mu_i^{(k)}(t_l),\nu_j^{(k)}(t_l),\Phi_j^{(k)}(t_l),s_j^{(k)}(t_l),\theta^{(k)}(t_l)\right)\right\}_{l=1}^{L}$,
        and truncate any leftover sample from the $k$-th
        run after
        $\left(\delta_i^{(k)}(t_L),\mu_i^{(k)}(t_L),\nu_j^{(k)}(t_L),\Phi_j^{(k)}(t_L),s_j^{(k)}(t_L),\theta^{(k)}(t_L)\right)$.
    end if
    Compute rank statistic for $i=1,\dots,q,j=1,\dots,n$:
      $r^{(k)}[\delta_i] = r\left(\left\{\delta_i^{(k)}(t_1),\dots,\delta_i^{(k)}(t_L)\right\},\widetilde{\delta_i^{(k)}}\right)$          (10)
                 $= \sum_{l=1}^{L} \mathbb{1}_{\left\{t_l:\delta_i^{(k)}(t_l)<\widetilde{\delta_i^{(k)}}\right\}}\left(\delta_i^{(k)}(t_l)\right)$.          (11)
    Similarly, calculate $r^{(k)}[\mu_i]$, $r^{(k)}[\nu_j]$, $r^{(k)}[\Phi_j]$, $r^{(k)}[s_j]$,
    $r^{(k)}[\theta]$.
end while
Plot the histogram of rank statistic $r_{ij}^{(k)}$. for $k=1,\dots,K$.
Check the uniformity of the histogram of $r_{ij}^{(k)}$. for
$k=1,\dots,K$.
```

This illustrates the applicability of the techniques in [7] for diagnosing the estimation of parameters in a BHM such as that in Figure 9.

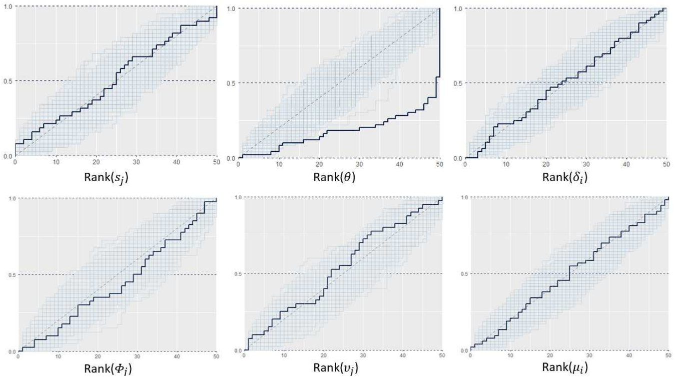

Fig. 7: SBC results of non-regression BASiCS [4]. For the model parameters s_j, δ_i, θ, Φ_j, ν_j and μ_i, the ECDF of the calculated rank statistic (dark blue) and 500 uniform samples (light blue) are plotted. Without loosing generality, here $i = 1$, $j = 1$.

We also perform the Simulation-based Calibration procedure described in Algorithm 1 on regression BASiCS model, adapted from [7]. Similar to the arguments in the last paragraph, without losing generality, in Figure 8 we plot the ECDF for the calculated rank statistics and the uniform distribution for δ_1, s_1 and θ. In terms of the Simulation-based Calibration results, the behaviours observed for the regression BASiCS model in Figure 8 are similar to those observed for the non-regression BASiCS model in Figure 7. In particular, the rank statistics for most parameters in Figure 8 are close to a uniform distribution. The low ranks of s_j are seen slightly more often in the computed ranks than we would expect from a uniform distribution, and the rank statistic of θ is far from the range of the uniform distribution. Thus, this suggests that θ tends to be underestimated in the regression BASiCS model.

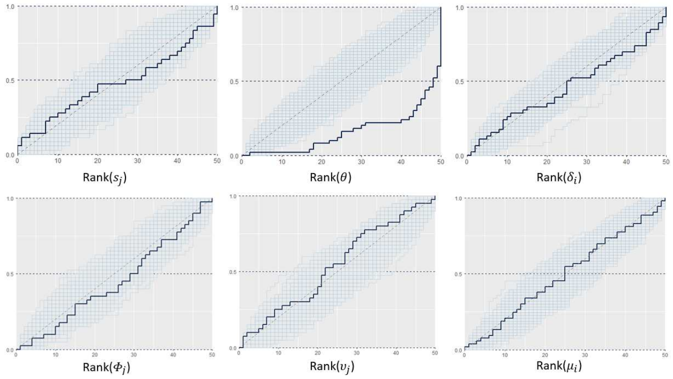

Fig. 8: SBC results of regression BASiCS [5]. For the model parameters s_j, δ_i and θ, Φ_j, ν_j, μ_i, the ECDF of the calculated rank statistic (dark blue) and 500 uniform samples (light blue) are plotted. Without loosing generality, here $i = 1$, $j = 1$.

IV. Conclusion

In summary, we have demonstrated how to exploit a simulation-based calibration approach to evaluate a BHM and its implementation. Our analysis shows that regression-BASiCS achieves some improvement over non-regression BASiCS on the posterior estimation accuracy in terms of the length of 89% credible interval and posterior predictive distribution. The Simulation-based Calibration method returns similar results for the two models. This is because the Simulation-based Calibration approach implemented here relies on checking if the true value of the parameter used to generate the corresponding synthetic dataset falls inside the posterior credible interval estimated under the assumed model [7]. In terms of the posterior credible interval coverage, these SBC results in Section III.D are consistent with the previous analysis about credible intervals in Section III.A. On the other hand, SBC results in Section III.D reveal issues with the posterior inference for some parameters, especially θ, which could affect the downstream analysis when implementing these models with real data. Considering that the prior distribution of θ is the same in both the regression BASiCS and the non-regression BASiCS, some improvements on this part of the model could be made in the future. We note that, in our BASiCS example, the point estimate section suggests a reliability issue with the posterior estimate of δ_i. Therefore, in future work on BHM, more point estimate options need to be explored. Since ground truth is typically unknown in BHM, we would like to emphasise that the simulation based reliability analysis is important in validating BHM and its implementation.

References

[1] P. Congdon, *Applied bayesian modelling*. John Wiley & Sons, 2014, vol. 595.

[2] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, K. Lao, and M. A. Surani, “mrna-seq whole-transcriptome analysis of a single cell,” *Nature Methods*, vol. 6, p. 377–382, 2009.

[3] C. A. Vallejos, J. C. Marioni, and S. Richardson, “BASiCS: Bayesian analysis of single-cell sequencing data,” *PLOS Comput. Biol.*, vol. 11, no. 6, p. e1004333, 2015.

[4] C. A. Vallejos, S. Richardson, and J. C. Marioni, “Beyond comparisons of means: understanding changes in gene expression at the single-cell level,” *Genome biology*, vol. 17, no. 1, pp. 1–14, 2016.

[5] N. Eling, A. C. Richard, S. Richardson, J. C. Marioni, and C. A. Vallejos, “Correcting the mean-variance dependency for differential variability testing using single-cell rna sequencing data,” *Cell Systems*, vol. 7, pp. 284–294, 2018.

[6] P. Brennecke, S. Anders, J. K. Kim, A. A. Kolodziejczyk, X. Zhang, V. Proserpio, B. Baying, V. Benes, S. A. Teichmann, J. C. Marioni, and M. G. Heisler, “Accounting for technical noise in single-cell RNA-seq experiments,” *Nature Methods*, vol. 10, pp. 1093–1095, 2013.

[7] S. Talts, M. Betancourt, D. Simpson, A. Vehtari, and A. Gelman, “Validating bayesian inference algorithms with simulation-based calibration,” *arXiv preprint arXiv:1804.06788*, 2018.

[8] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/

[9] D. Makowski, M. S. Ben-Shachar, and D. Lüdecke, “bayestestr: Describing effects and their uncertainty, existence and significance within the bayesian framework.” *Journal of Open Source Software*, vol. 4, no. 40, p. 1541, 2019. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.01541

[10] A. Gelman and C. R. Shalizi, "Philosophy and the practice of bayesian statistics," *British Journal of Mathematical and Statistical Psychology*, vol. 66, no. 1, pp. 8–38, 2013.

[11] D. J. Schad, M. Betancourt, and S. Vasishth, "Toward a principled bayesian workflow in cognitive science." *Psychological methods*, vol. 26, no. 1, p. 103, 2021.

[12] Statisticat and LLC., *LaplacesDemon: Complete Environment for Bayesian Inference*, 2021, r package version 16.1.6. [Online]. Available: https://web.archive.org/web/20150206004624/http://www.bayesian-inference.com/software

[13] A. S. Devonshire, R. Elaswarapu, and C. A. Foy, "Evaluation of external rna controls for the standardisation of gene expression biomarker measurements," *BMC Genomics*, vol. 11, no. 662, 2010.

[14] M. Greenwood and G. U. Yule, "An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference of multiple attacks of disease or of repeated accidents," *Journal of the Royal Statistical Society*, vol. 83, no. 2, pp. 255–279, 1920.

[15] J. R. Peat, W. Dean, S. J. Clark, F. Krueger, S. A. Smallwood, G. Ficz, J. K. Kim, J. C. Marioni, T. A. Hore, and W. Reik, "Genome-wide bisulfite sequencing in zygotes identifies demethylation targets and maps the contribution of tet3 oxidation," *Cell reports*, vol. 9, no. 6, pp. 1990–2000, 2014.

Appendix

A. The Non-regression BASiCS Model

BASiCS [3]–[5] aims to provide a structural method to analyse scRNAseq count data to detect highly variable genes (HVGs) and lowly variable genes (LVGs). HVGs are expressed differently across cells because they specialise in certain functions of certain cells. On the opposite, LVGs are expressed on a stable level across cells, as they participate in general cellular activities.

A typical scRNA dataset can be represented by a matrix, $\mathbf{X} = \left(X_{ij}\right)_{i\in\{1,\ldots,q\},j\in\{1,\ldots,n\}}$ where non-negative integers $X_{ij} \in \{0,1,2,\ldots\}$ represent the mRNA count of gene i in cell j at the time of the experiment, for q genes and n cells. Here we consider q_0 *biological* genes, which naturally exist within the cells, and $q-q_0$ *spike-in* genes [13], which are artificially added during the experiment to help quantify the technical noise.

The expected count of gene i's expression in cell j could be affected by several factors. BASiCS [3]–[5] provides a hierarchical statistical model for these gene expression counts, mainly based on Assumptions 1-5 below. A schematic representation of this hierarchical model is given in Figure 9.

a) Assumption 1: The expression count for biological gene $i \in \{1,...,q_0\}$ in cell $j \in \{1,...,n\}$ can be modelled as

$$X_{ij}\left|\mu_i, \Phi_j, \nu_j, \rho_{ij} \overset{ind}{\sim} \text{Poisson}(\mu_i \Phi_j \nu_j \rho_{ij}).\right. \tag{12}$$

On the other hand, the expression count for spike-in gene $i \in \{q_0+1,...,q\}$ in cell $j \in \{1,...,n\}$ is modelled as

$$X_{ij}\left|\mu_i, \nu_j \overset{ind}{\sim} \text{Poisson}(\mu_i \nu_j).\right. \tag{13}$$

Here, an underlying assumption is that the unexplained technical noise only depends on cell-specific characteristics. For a given cell j, this noise would affect the expression counts of all genes $i = 1,...,q$ in the same manner.

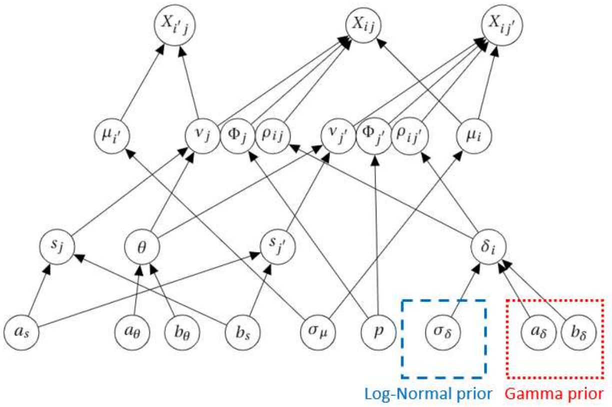

Fig. 9: A schematic representation of the non-regression BASiCS model in terms of a directed acyclic graph. The two choices for the prior distribution for δ_i, log-Normal and Gamma distributions, are depicted. In the graph, we have biological genes $i \in \{1,\ldots,q_0\}$, spike-in genes $i' \in \{q_0+1,\ldots,q\}$ and cells $j, j' \in \{1,\ldots,n\}$.

b) Assumption 2: The biological random effect for gene $i \in \{1,\ldots,q_0\}$ in cell $j \in \{1,\ldots,n\}$ follows a Gamma distribution which depends on the gene but not on the cell,

$$\rho_{ij}|\delta_i \overset{ind}{\sim} \text{Gamma}\left(\frac{1}{\delta_i}, \frac{1}{\delta_i}\right). \tag{14}$$

Both the shape and the rate of the Gamma distribution are assumed to be the same, δ_i^{-1}, so that $\mathrm{E}(\rho_{ij}) = 1$, $\mathrm{Var}(\rho_{ij}) = \delta_i$. The biological variation factor for biological gene $i \in \{1,\ldots,q_0\}$ across all cells, δ_i, has two possible options for prior [3], [4]:

$$\textbf{Log-normal prior:} \qquad \delta_i|\sigma_\delta \overset{ind}{\sim} \text{log-Normal}\left(0, \sigma_\delta^2\right), \tag{15}$$

$$\textbf{Gamma prior:} \qquad \delta_i|a_\delta, b_\delta \overset{ind}{\sim} \text{Gamma}\left(a_\delta, b_\delta\right), \tag{16}$$

with the corresponding standard deviation, shape and rate parameters $\sigma_\delta, a_\delta, b_\delta > 0$.

c) Assumption 3: The expected expression count of biological gene $i \in \{1,\ldots,q_0\}$, μ_i, follows a log-Normal distribution,

$$\mu_i\left|\sigma_\mu \overset{ind}{\sim} \text{log-Normal}\left(0, \sigma_\mu^2\right), \quad \text{with } \sigma_\mu > 0.\right. \tag{17}$$

d) Assumption 4: The cell size variables follow a scaled Dirichlet distribution,

$$(\Phi_1,...,\Phi_n)\left|\boldsymbol{p} \sim n\text{Dirichlet}(\boldsymbol{p}),\right. \tag{18}$$

where $\boldsymbol{p} = \left(p_1,...,p_n\right)$ is the concentration parameter, with $p_1,\ldots,p_n > 0$. The Dirichlet prior also restricts that $n\left(\sum_{j=1}^n \Phi_j\right)^{-1} = 1$.

e) Assumption 5: The technical variation factor ν_j, for cell $j \in \{1,\ldots,n\}$, follows a Gamma distribution

$$\nu_j\left|\theta, s_j \overset{ind}{\sim} \text{Gamma}\left(\frac{1}{\theta}, \frac{1}{s_j\theta}\right),\right. \tag{19}$$

with shape parameter $\frac{1}{\theta} > 0$ and rate parameter $\frac{1}{s_j\theta} > 0$, so that $\mathrm{E}(\nu_j) = s_j$, $\mathrm{Var}(\nu_j) = s_j^2\theta$. The general technical noise factor across all cells, θ, and the technical noise related to specific cell $j \in \{1, \ldots, n\}$, are modelled as

$$\theta \,|\, a_\theta, b_\theta \sim \text{Gamma}(a_\theta, b_\theta), \tag{20}$$

$$s_j \,|\, a_s, b_s \overset{i.i.d.}{\sim} \text{Gamma}(a_s, b_s), \tag{21}$$

with shape and rate parameters $a_\theta, b_\theta, a_s, b_s > 0$.

1) Identifiability of the Non-regression BASiCS: Considering the distributions given by Eqs. (12) and (14), We apply the Poisson-Gamma mixture result in [14]. In particular, after integrating out ρ_{ij} for $i = 1, \ldots, q_0$, we get the likelihood of the gene expression count of biological genes and spiked-in genes:

$$X_{ij} \,\middle|\, \mu_i, \delta_i, \Phi_j, \nu_j \overset{ind}{\sim} \begin{cases} \text{Neg-Binomial}\left(\delta_i^{-1}, \frac{\Phi_j \nu_j \mu_i}{\Phi_j \nu_j \mu_i + \delta_i^{-1}}\right), \\ \qquad\qquad \textit{for } i \in \{1, ..., q_0\}, \\ \text{Poisson}\left(\nu_j \mu_i\right), \\ \qquad\qquad \textit{for } i \in \{q_0 + 1, ..., q\}, \end{cases} \tag{22}$$

for $j \in \{1, \ldots, n\}$. The resulting model in Eq. (22) for biological genes is not identifiable, in the sense that it is to be expected that parameters μ_i, ν_j and Φ_j cannot be separately estimated from gene expression data for biological genes, since they appear multiplied as $\mu_i \nu_j \Phi_j$ in the expression above. However, spike-in genes facilitate identifiability.

For spiked-in genes, $i \in \{q_0 + 1, \ldots, q\}$, the expected count μ_i is known, since the number of spiked-in molecules added to each cell is recorded. Therefore, using the spiked-in information across the cells $j \in \{1, \ldots, n\}$, the posterior distribution of ν_j can be inferred from $X_{ij} \sim \text{Poisson}\left(\nu_j \mu_i\right)$, $i \in \{q_0 + 1, \ldots, q\}, j \in \{1, \ldots, n\}$. In particular,

$$\begin{aligned} p\left(\nu_j | X_{ij}, \mu_i\right) &\propto \prod_{i=q_0+1}^{q} p(X_{ij} | \nu_j, \mu_i) p(\nu_j) \\ &= \frac{\left(\frac{1}{\theta}\right)^{\frac{1}{s_j \theta}} \prod_{i=q_0+1}^{q} \mu_i^{X_{ij}}}{X_{ij}! \Gamma\left(\frac{1}{\theta}\right)} \nu_j^{\left[\sum_{i=q_0+1}^{q} X_{ij} + \frac{1}{\theta} - 1\right]} e^{-\left(\mu_i + \frac{1}{s_j \theta}\right) \nu_j}, \end{aligned}$$

where μ_i and X_{ij} are known for $i \in \{q_0+1, \ldots, q\}$, $j \in \{1, \ldots, n\}$. Since ν_j for $j \in \{1, \ldots, n\}$ are inferred from spiked-in information, the remaining identifiability conflict is between Φ_j, $j \in \{1, \ldots, n\}$, and the expected counts μ_i, $i \in \{1, \ldots, q_0\}$. However, the restriction $n\left(\sum_{j=1}^{n} \Phi_j\right)^{-1} = 1$ in the Dirichlet distribution ensures the identifiability of Φ_j, $j \in \{1, \ldots, n\}$ and μ_i, $i \in \{1, \ldots, q_0\}$. We note that this restriction imposes an arbitrary scale to Φ_j, but it does not affect the relative differences between the μ_i's nor the δ_i's.

B. The Regression BASiCS Model

According to [6], a strong relationship is typically observed between the variability (δ_i in BASiCS [5]) and mean (μ_i in BASiCS [5]) estimates. In this case, the interpretation of results from the non-regression BASiCS model can be hindered. As authors in [5] argued, an intuitive approach would be to only compare variability δ_i of those genes with equal mean expression μ_i, but this is sub-optimal, especially when used between groups of cells, as there are a large number of genes expressed differently between populations. One such example is provided by [5], where reactive genes that change in mean expression upon changing conditions are excluded from the expression heterogeneity assessment by this intuitive solution. An alternative solution is to directly adjust variability measures to remove the confounding effect between mean and variability. For example, [15] computed the empirical distance between the squared coefficient of variation (CV^2) of each gene to the rolling median CV^2 across genes with similar expression levels. In line with this approach, authors in [5] introduce the joint prior distribution for the expected expression count of gene i, μ_i, and the gene-specific hyper-parameter δ_i:

$$\mu_i \,|\, \sigma_\mu \overset{ind}{\sim} \text{log-Normal}\left(0, \sigma_\mu^2\right), \tag{23}$$

$$\delta_i \,|\, \mu_i, \beta, \sigma_\delta^2, \lambda_i \overset{ind}{\sim} \text{log-Normal}\left(f(\mu_i), \frac{\sigma_\delta^2}{\lambda_i}\right), \tag{24}$$

for $i \in \{1, \ldots, q_0\}$. The latter is equivalent to consider the non-linear regression model

$$\log(\delta_i) = f(\mu_i) + \omega_i, \quad i \in \{1, ..., q_0\}, \tag{25}$$

where $\omega_i \,|\, \sigma_\delta^2, \lambda_i \overset{ind}{\sim} \text{Normal}\left(0, \sigma_\delta^2 \lambda_i^{-1}\right)$ is a latent gene-specific residual over-dispersion parameter, capturing departures from the overall trend across all genes expressed at a given mean expression μ_i. The regression BASiCS considers

$$f(\mu_i) = \alpha_0 + \alpha_1 \log\left(\mu_i\right) + \sum_{l=1}^{L} \beta_l g_l \left(\log\left(\mu_i\right)\right), \quad i \in \{1, \ldots, q_0\} \tag{26}$$

where $\alpha_0, \alpha_1, \beta_1, ..., \beta_L$ are regression coefficients and $g_1(\cdot), ..., g_L(\cdot)$ represent a set of Gaussian Radial Basis Function (GRBF) kernels. These can be defined as

$$g_l\left(\log\left(\mu_i\right)\right) = \exp\left\{-\frac{1}{2}\left(\frac{\log\left(\mu_i\right) - m_i}{h_l}\right)^2\right\}, \quad l \in \{1, \ldots, L\} \tag{27}$$

where m_l, $l \in \{1, \ldots, L\}$, are location hyperparameters and h_l, $l \in \{1, \ldots, L\}$, are scale hyperparameters. Therefore, on top of Assumptions 1-5, one adds Assumption 6 for regression BASiCS [5], leading to the HBM given by Eqs. (18)-(30).

a) Assumption 6: A priori, m_l ($l \in \{1, ..., L\}$), h_l ($l \in \{1, ..., L\}$) and σ_δ^2 are fixed. The priors for $\beta = \left(\alpha_0, \alpha_1, \beta_1, \ldots, \beta_L\right)$ in Eq. (16) and its hyperparameters are proposed as below:

$$\beta | \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}\right) \tag{28}$$

$$\sigma^2 \overset{ind}{\sim} \text{Inv-Gamma}\left(a_\sigma, b_\sigma\right) \tag{29}$$

$$\lambda_i | \eta \overset{ind}{\sim} \text{Gamma}\left(\frac{\eta}{2}, \frac{\eta}{2}\right), \; i \in \{1, \ldots, q_0\}. \tag{30}$$