Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation

Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation

Elena Sabbioni[a], Enrico Bibbona[a], Gianluca Mastrantonio[a], and Guido Sanguinetti[b]

[a]Politecnico di Torino; elena.sabbioni@polito.it, enrico.bibbona@polito.it, gianluca.mastrantonio@polito.it,
[b]Scuola Internazionale Superiore di Studi Avanzati (SISSA); gsanguin@sissa.it

Abstract

The analysis of RNA data plays a crucial role in understanding cellular differentiation. One widely-used methodology for analyzing RNA data is scVelo. However, in this paper, we show that, among other issues of scVelo, the current model formalization suffers from identifiability problems. We propose a Bayesian version of scVelo with modifications that address these issues.

Keywords: scVelo, RNA, Bayesian, identifiability

1. Introduction

RNA velocity is a critical biological metric that facilitates the reconstruction of cellular differentiation at single-cell level. It provides insight into the future state of each cell, and it is closely associated with transcription from DNA to RNA, as well as the quantity of spliced mRNA in each cell. By analyzing RNA velocity, researchers can gain a deeper understanding of the underlying mechanisms driving cellular differentiation, which has important implications for fields such as developmental biology and disease research. Single-cell RNA sequencing (scRNA-seq) techniques are commonly used to measure the abundance of unspliced and spliced mRNA in each cell for each gene, which is essential for inferring RNA velocity. However, these techniques are destructive, as they permit only a single observation of gene expression for each cell before it is destroyed. In this sector, one of the most influential works is scVelo, presented in [1]. Despite its success in the scientific community, there are several criticisms when it is analyzed from a mathematical and statistical point of view.

As a primary contribution, we reframe the model under a Bayesian framework, which provides better insight into the parameters that can be estimated and identified. The use of Bayesian inference enables us to compare the posterior estimates of the parameters with their corresponding priors and compute credible intervals. In contrast, such comparisons and interval estimates are not possible with the point estimates provided by scVelo. Through simulated examples, we demonstrate that the "time" parameter is not identifiable, which, to the best of our knowledge, has not been previously identified in the literature, representing another contribution of this paper. Furthermore, we propose modifications to the model that addresses some of the criticisms of the original scVelo. Collectively, these contributions have the potential to enhance the accuracy and reliability of RNA velocity inference, as well as provide new insights into cellular differentiation.

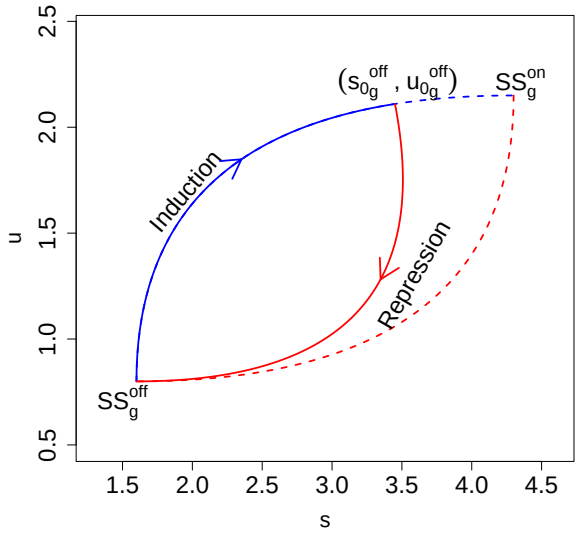


Figure 1: Solution of (2) in the space (s, u) for gene g. The solid line represents the gene's behavior if an early switch at time t_{0g}^{off} occurs, while the dashed line represents the potential behavior if the gene reaches the steady state SS_g^{on} during the inductive phase. The blue upper curve corresponds to the gene's inductive phase, characterized by the rate α_g^{on}, and the red lower curve represents the repressive phase, associated with α_g^{off}.

2. Mathematical model

Let us consider a system with n_g genes and n_c cells. The model assumes a straightforward chemical reaction network (CRN) to represent the processes of transcription, splicing and degradation, with gene-specific rates, according to mass-action kinetics. The CRN representation of the process, for a given g, is

$$\text{DNA} \xrightarrow{\alpha_g^k} \text{unspliced mRNA} \xrightarrow{\beta_g} \text{spliced mRNA} \xrightarrow{\gamma_g} \emptyset \tag{1}$$

that is associated with the following ordinary differential equation (ODE) system:

$$\begin{cases} \frac{du_g(t)}{dt} = \alpha_g^k - \beta_g u_g(t), \\ \frac{ds_g(t)}{dt} = \beta_g u_g(t) - \gamma_g s_g(t), \\ u_g\left(t_{0g}^k\right) = u_{0g}^k, \\ s_g\left(t_{0g}^k\right) = s_{0g}^k, \end{cases} \tag{2}$$

where $u_g(t)$ and $s_g(t)$ are the reads of unspliced and spliced mRNA at time t in a cell. The solution to the ODEs is

$$\begin{cases} u_g(t) = u_{0g}^k e^{-\beta_g \tau_g^k} + \frac{\alpha_g^k}{\beta_g}(1 - e^{-\beta_g \tau_g^k}) \\ s_g(t) = s_{0g}^k e^{-\gamma_g \tau_g^k} + \frac{\alpha_g^k}{\gamma_g}(1 - e^{-\gamma_g \tau_g^k}) + \frac{\alpha_g^k - \beta_g u_{0g}^k}{\gamma_g - \beta_g}(e^{-\gamma_g \tau_g^k} - e^{-\beta_g \tau_g^k}) \quad \text{with } \tau_g^k = t - t_{0g}^k. \end{cases} \tag{3}$$

It should be noted that the variable t is not the real time, but a representation of the cell position in the ODE dynamic, which is often called *pseudotime* in the literature, see, for example, [3].

RNA velocity is defined as

$$v_g(t) := \frac{ds_g(t)}{dt} = \beta_g u_g(t) - \gamma_g s_g(t).$$

Accurately estimating the model parameters is crucial for obtaining a reliable estimator of this biological quantity.

Parameters description For each gene, there exist two transcription rates, indicated as α_g^{on} and α_g^{off} with $\alpha_g^{\text{on}} > \alpha_g^{\text{off}}$, represented in (2) as α_g^k, with $k \in \{\text{on}, \text{off}\}$. The rates regulate the conversion of DNA into unspliced mRNA, as depicted in (1) (first and second components). This implies that a gene can exist in two distinct states: an inductive phase, regulated by the transcription rate α_g^{on}, and a repressive phase, where transcription either occurs at a lower rate or is absent altogether, dictated by α_g^{off}. It is assumed that each gene can be activated and then repressed only once, which is justified by the assumption that the total time length of the biological processes is sufficiently small. The rates β_g and γ_g, illustrated in (1) (from the second to the fourth component), are responsible for the splicing and degradation mechanisms of mRNA. The gene time dynamic is depicted in Figure 1.

The ODE system has two theoretical steady states, that are only gene-dependent and are identified by the coordinates

$$\text{SS}_g^{\text{off}} = \left(\frac{\alpha_g^{\text{off}}}{\beta_g}, \frac{\alpha_g^{\text{off}}}{\gamma_g} \right), \qquad \text{SS}_g^{\text{on}} = \left(\frac{\alpha_g^{\text{on}}}{\beta_g}, \frac{\alpha_g^{\text{on}}}{\gamma_g} \right),$$

in the space (s, u), see Figure 1. After the cell is created, each gene remains at SS_g^{off} for a time period of t_{0g}^{on} before being activated and entering the inductive phase, represented by the upper blue arc in Figure 1. Time t_{0g}^{on} is not identifiable (see Section 3) since there is no information in the data regarding the real time point at which the cells are observed, and hence, without loss of generality, we assume $t_{0g}^{\text{on}} = 0$. This means that $s_{0g}^{\text{on}} = \alpha_g^{\text{off}}/\beta_g$ and $u_{0g}^{\text{on}} = \alpha_g^{\text{off}}/\gamma_g$. However, before reaching the second steady state SS_g^{on}, the repressive phase is triggered at time $t_{0g}^{\text{off}} + t_{0g}^{\text{on}}$ and the dynamic follows the evolution depicted by the solid line in Figure 1.

In scVelo the pre-processed data $(Y_{u,cg}, Y_{s,cg})'$ are assumed to be normally distributed and the unspliced and spliced components to be independent, i.e.

$$Y_{u,cg} \sim \mathcal{N}(u_g(t_{cg}), \sigma^2) \qquad Y_{s,cg} \sim \mathcal{N}(s_g(t_{cg}), \sigma^2) \qquad Y_{u,cg} \perp\!\!\!\perp Y_{s,cg}. \tag{4}$$

where $t_{cg} = \tau_{cg} + t_{0g}^{k_{cg}}$ and $u_g(t_{cg})$ and $s_g(t_{cg})$ are evaluated in a time that is both cell- and gene-specific. The description of the data used to estimate the model is discussed in Section 3. The scVelo algorithm estimates the following parameters: $(\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \beta_g, \gamma_g, t_{0g}^{\text{off}})$ for each gene, and (τ_{cg}, k_{cg}) for each cell and gene.

There are several criticisms of this model, that, in our opinion, raise questions about the reliability of the results, which will be discussed in the next section.

3. Critical issues of scVelo

One of the main concerns is related to the estimation of the cell- and gene-specific τ_{cg}. Since single-cell data only provides a single observation for each cell, inferring τ_{cg} is inherently difficult, if not impossible. Despite this, the authors did not acknowledge this issue. As a first contribution, we demonstrate in Section 4. that τ_{cg} is at best weakly identifiable by showing that, the posterior distribution of τ_{cg} closely resembles the one of $\tau_{c'g}$, for $c \neq c'$, and they are both very similar to the prior.

Single-cell RNA-sequencing dataset contains discrete counts, describing the number of measured RNA molecules in each cell. In scVelo, a series of pre-processing steps are applied to the raw data. This includes filtering out genes that are not expressed, normalizing the counts to account for differences in sequencing depth across cells, and smoothing the gene expression profiles among groups of cells with similar genetic expressions. Additionally, the logarithm of the pre-processed counts is taken. The variables $(Y_{u,cg}, Y_{s,cg})'$ in equation (4), are the results of this pre-processing. While these pre-processing steps are common in many biological pipelines, they significantly alter the nature of the data by transforming them from discrete counts to continuous values. This transformation can be problematic, especially in real data applications where the original counts are often very low (often in the range [0,10]). The pipeline introduces dependence across genes which are not taken into account in the model, that assumes independence, see (2) and (4). Additionally, the use of a logarithmic transformation is questionable since ODE equations and solutions are not invariant under a non-linear transformation.

Despite time-dynamic being dependent on four parameters $(\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \beta_g, \gamma_g)'$, only three of them are identifiable due to the lack of information on t_{cg} and its scale in the data. We can easily see that, if $r \in \mathbb{R}^+$, then the parameters $(\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \beta_g, \gamma_g)'$ and $(\alpha_g^{\text{off}}/r, \alpha_g^{\text{on}}/r, \beta_g/r, \gamma_g/r)'$ produce the same likelihood if τ_{cg} is substituted with $r\tau_{cg}$. This is because under both sets of parameters, the same value $(u_g(t_{cg}), s_g(t_{cg}))$ is obtained. While some of the issues discussed here have been previously addressed in the literature (e.g., [2; 5]), the non-identifiability of τ_{cg} and its impact on other parameter estimates has not been adequately emphasized to the scientific community.

In conclusion, scVelo has a further drawback in that it only provides point estimates of the parameters and does not compute any measure of their precision. This absence reduces the reliability of the results as it is not possible to assess the statistical differences among the parameters accurately.

4. The Bayesian Implementation

In our model formulation, we choose to use the original data without the non-linear pre-processing steps applied in scVelo. The only step we keep is the filtration of the genes that are not sufficiently expressed. As a result, our data $(Y_{u,cg}, Y_{s,cg})'$ is discrete, and $(u_g(t_{cg}), s_g(t_{cg}))'$, obtained as solution of (2), represents the mean of the original count data. A natural choice for modeling $(Y_{u,cg}, Y_{s,cg})'$ is the Poisson, because this distribution arises from the chemical master equation (CME) associated with the CRN (1). Specifically, in the steady state, the Poisson distribution is the distribution of mRNA counts of a single gene in a single cell, and in the transient part, CME distribution can be expressed as the convolution of multinomial and product Poisson distributions [4]. On the other hand, to increase the model flexibility and to take into account extra sources of variability we use Negative Binomial distribution, which is an overdispersed version of the Poisson. Specifically, we assume that

$$Y_{u,cg} \sim \mathcal{NB}(u_g(t_{cg}), \eta_g) \qquad Y_{s,cg} \sim \mathcal{NB}(s_g(t_{cg}), \eta_g) \qquad Y_{u,cg} \perp\!\!\!\perp Y_{s,cg}$$

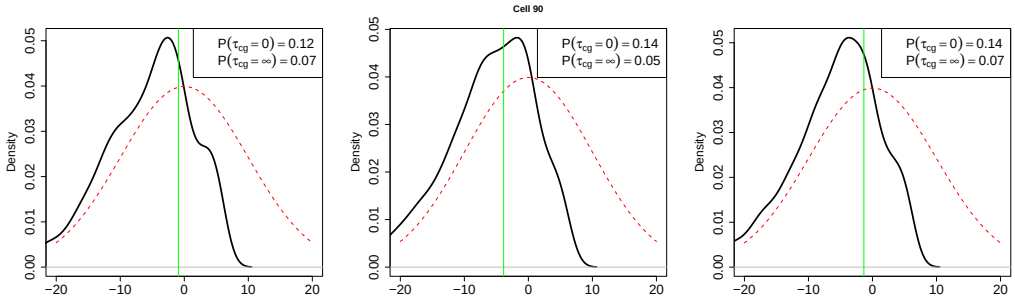
Here the Negative-Binomial is parameterized in terms of its mean μ and the overdispersion parameter η, such that if $X \sim \mathcal{NB}(\mu, \eta)$, then $\mathbb{V}(X) = \mu(1 + \mu\eta)$. As prior distributions we define $\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \gamma_g, \eta_g \sim \mathcal{N}_{[0,+\infty)}(0, 10000)$, where $\mathcal{N}_{[a,b]}$ is a truncated Normal distribution with support in $[a, b]$, and $P(k_{cg} = \text{on}) = 0.5$. For τ_{cg} we define a mixed-type distribution with two masses of value 0.1 on $\tau_{cg} = 0$ and $\tau_{cg} = \infty$, respectively, and with probability 0.8 we have $\log \tau_{cg} \sim N(0, 100)$. The two masses are used to locate the cell in the steady states. The prior on t_{0g}^{off} must depend on the set $\{\tau_{cg}, k_{cg}\}_{c=1}^{n_c}$ since

$$t_{0g}^{\text{off}} \geq \max\{\tau_{cg} | k_{cg} = \text{on}, c = 1, \ldots, n_c\} = \tau_{g,\max}^{\text{on}} \tag{5}$$

hence, we assume the following: $\log t_{0g}^{\text{off}} | \{\tau_{cg}, k_{cg}\}_{c=1}^{n_c} \sim N_{[\log \tau_{g,\max}^{\text{on}}, \infty)}(0, 100)$ To avoid identifiability issue, parameter β_g is fixed to 1.

Simulation setting We simulate data with $n_g = 5$, $n_c = 3600$, and γ_g randomly generated in $[0.5, 0.8]$, α_g^{off} in $[0.05, 1]$, α_g^{on} in $[2, 5]$, and η_g in $[0.01, 0.1]$. These intervals have been chosen such that the empirical distribution of the raw data mimics the one of the real pancreatic dataset used in [1]. There are different issues when simulating parameters τ_{cg}, t_{0g}^{off}, and k_{cg}. Indeed, we have to satisfy equation (5) and, to have realistic and diversified locations of $(s_g(t_{cg}), u_g(t_{cg}))$, as well as $(s_g(t_{0g}^{\text{off}}), u_g(t_{0g}^{\text{off}}))$ close to SS_g^{on}, SS_g^{off} or in between. To achieve this, a sequence of if/else conditions were implemented, however for the sake of brevity these details are omitted. We run the model for 100000 iterations, with thin 40 and burning 10000, having then 2250 posterior samples.

Discussion of the results The posterior distributions of different τ_{cg}, as shown in Figure 2, reveal that there are minimal differences between them. Additionally, the posterior distribution is similar to the prior distribution. Comparable results are obtained when changing the prior distribution, which are not presented for the sake of brevity. It should be pointed out that, under this setting, for $\log \tau_{cg} < -5$ and $\log \tau_{cg} > 5$ the coordinates $(s_g(t_{cg}), s_g(t_{cg}))$, for all $c = 1, \ldots, n_c$ and $g = 1, \ldots, n_g$, are approximately equal to the steady states with a difference of order 10^{-3}.



(a) (b) (c)

Figure 2: Gene- and cell-specific τ_{cg} model. Prior (red dashed line) and posterior (black solid line) of the logarithm of τ_{cg} for three cells. The vertical line represents the true value.

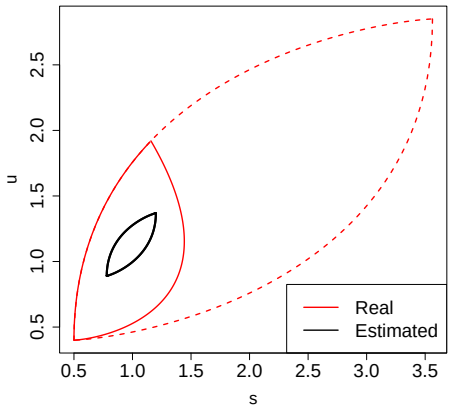


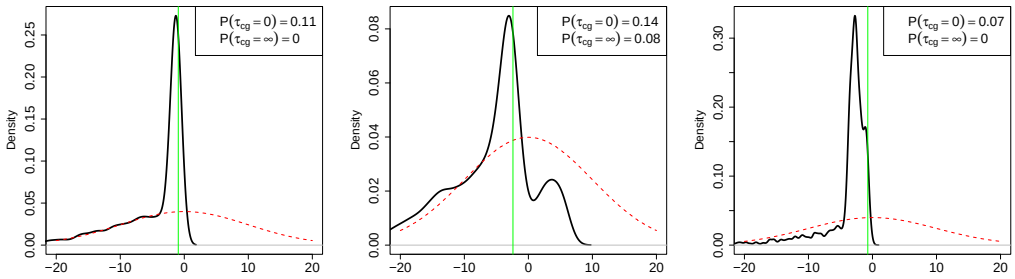
Figure 3: Phase plot in the space (s, u) for a given gene. The solid lines show the solutions (3) obtained with the real parameters (red) used to simulate the data and the posterior means (black). The dashed lines correspond to the potential dynamic if the steady state $\mathrm{SS}_g^{\mathrm{on}}$ is reached. For the estimated solution, the dashed and the solid line coincide.

This illustrates the difficulty in estimating these parameters and emphasizes the importance of considering the full distribution rather than just point estimates. This issue cannot be detected with the original scVelo implementation, which only provides point estimates as output. As a consequence of this weak identifiability, all the other parameters are wrongly estimated, i.e., the associated 95% credible intervals do not contain the true value, and the entire structure in the space (s, u) describing the time dynamic, is very different from the real one, as shown in Figure 3.

Simulated examples demonstrate that estimating the parameter τ_{cg} and other unknowns in the model is feasible when we have repeated measures for each (c, g). The results obtained with $n_c = 8$, $n_g = 5$ and 450 repetitions are shown in Figure 4 as an example. However, in the case of single-cell data, it is not possible to have true repetitions since the variable t_{cg} is unobservable/unknown. Instead of repetitions, a mixture model can be used where data share a common coordinate in the space (s, u). These results suggest that this approach may be a viable direction.

5. Conclusions and further developments

This study highlights the issues present in the current formalization of the scVelo model. Specifically, we focus on the weak identifiability of the variable τ_{cg} and its impact on the estimation of other parame-



(a) (b) (c)

Figure 4: Repeated measurements model. Prior (red dashed line) and posterior (black solid line) of the logarithm of τ_{cg} for a three genes. The vertical lines represent the true values.

ters. To address this, we introduce a new Bayesian version of scVelo and evaluate its performance using synthetic data. Upon inspection of the posterior distribution of τ_{cg}, we observe that, for a given gene, all distributions are comparable and closely resemble the prior distribution, indicating weak identifiability of the parameters.

In addition, we propose a potential solution to overcome the identifiability problem, which shows promising results in our initial investigations. We are pursuing this direction as a possible way forward in improving the performance of scVelo.

References

[1] Bergen, V. and Lange, M. and Peidli, S. and Wolf, F. A. and Theis, F. J.: Generalizing RNA velocity to transient cell states through dynamical modeling. Nature biotechnology **38.12**, 1408–1414 (2020)

[2] Gorin, G. and Fang, M. and Chari, T. and Pachter, L.: RNA velocity unraveled. PLOS Computational Biology **18.9** (2022)

[3] Gupta, R., Cerletti, D.Gut, G., Oxenius, A., Claassen, M.: Simulation-based inference of differentiation trajectories from RNA velocity fields. Cell Reports Methods, **2** 1–15 (2022)

[4] Jahnke, T. and Huisinga, W.: Solving the chemical master equation for monomolecular reaction systems analytically. Journal of mathematical biology **54.1** 1–26 (2007)

[5] Marot-Lassauzaie, V. and Bouman, B. J. and Donaghy, F. D. and Demerdash, Y. and Essers, M. A. G. and Haghverdi, L.: Towards reliable quantification of cell state velocities. PLOS Computational Biology **18.9** (2022)