Extending Bayesian Modelling of RNA Velocity

Extending Bayesian modelling of RNA velocity

Elena Sabbioni[1], Enrico Bibbona[1] Gianluca Mastrantonio[1], and Guido Sanguinetti[2]

[1] Politecnico di Torino, Torino, Italy,
[2] Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy
elena.sabbioni@polito.it

Abstract. This paper introduces an alternative method for modeling RNA velocity within the Bayesian framework, employing zero-inflated distributions without the need for artificial preprocessing to handle RNA counts. Through a comparative analysis conducted on a real dataset, we illustrate the performance of our approach, showcasing outcomes comparable to those achieved with assumptions of Negative Binomial data on preprocessed observations. Our proposed model eliminates the requirement for arbitrary data filtering, thereby demonstrating its effectiveness in capturing the underlying biological dynamics.

Keywords: Zero-Inflated; RNA velocity; Negative Binomial; Bayesian

1 Introduction

RNA velocity plays a pivotal role in biology by aiding in understanding of cellular differentiation at the individual cell level. It offers predictive insight into the future state of cells and is intricately linked to the processes of transcription from DNA to RNA, as well as the levels of spliced mRNA within each cell. Through RNA velocity analysis, researchers can deepen their comprehension of the fundamental processes steering cellular differentiation, with significant implications for fields like developmental biology and disease research. Single-cell RNA sequencing (scRNA-seq) methods are widely employed to quantify unspliced and spliced mRNA levels per gene in each cell, a crucial step for estimating RNA velocity. Within this field, a notable cornerstone is the work of scVelo, presented in [1], which, despite its widespread acceptance in the scientific community, faces multiple criticisms, as detailed in [2], [3].

We focus on the initial processing phase utilized in various biological pipelines, where genes are screened based on factors such as low expression, minimal total counts across all cells, or insufficient variability. The processed data is then subjected to modeling, assuming a Poisson, Negative Binomial, or Normal distribution post specific data transformations. These procedural steps introduce a degree of subjectivity and carry the risk of eliminating potentially crucial genes. For instance, genes with low expression levels, yet pivotal for gene dynamics, may be excluded. This is particularly relevant when considering that the cells analyzed in the RNA velocity framework typically belong to early developmental stages, where low counts are commonplace.

To mitigate the arbitrariness in the selection step, we propose adopting a modeling approach. Rather than working with the filtered data, we use the original data, which often includes a substantial number of zero counts. In this approach, we enhance the standard modeling procedure by incorporating a two-stage modeling strategy. This involves jointly modeling the zeros and positive counts using a zero-inflated distribution based on the Poisson.

2 The model

We indicate with $(y_{s,cg}, y_{u,cg})$ the observed counts of spliced and unspliced mRNA in cell $c \in \{1, \dots n_c\}$ and gene $g \in \{1, \dots n_g\}$, where n_c and n_g denote the total number of cells and genes, respectively. In the biological literature, in handling the observed counts it is common to assume that all variables are independent and follow either a Poisson distribution

$$Y_{s,cg} \sim P(\mu_{s,cg}) \quad \text{and} \quad Y_{u,cg} \sim P(\mu_{u,cg}),$$

or a Negative Binomial (NB)

$$Y_{s,cg} \sim NB(\mu_{s,cg}, \eta_g) \quad \text{and} \quad Y_{u,cg} \sim NB(\mu_{u,cg}, \eta_g).$$

Here, $\mu_{\cdot,cg}$ represents the expected value that, in this context, is a function of the solution of an Ordinary Differential Equation (ODE) system, which will be introduced in Section 2.2. Additionally, η_g is the overdispersion parameter, such that, under the Negative Binomial assumption, we have

$$\mathbb{E}\left[Y_{\cdot,cg}\right] = \mu_{\cdot,cg} \quad \text{and} \quad \mathbb{V}ar\left(Y_{\cdot,cg}\right) = \mu_{\cdot,cg} + \mu_{\cdot,cg}^2 \eta_g.$$

As mentioned before, it is important to note that cells analyzed in the RNA velocity framework are typically in an early stage of development. Hence, many genes may not yet be expressed, leading to large proportion of zeros, which may not be likely under the previous data assumption.

This challenge, under these data distributions, is commonly addressed by excluding "non-expressed" genes. However, here we propose an alternative approach, generally referred to as a zero-inflated approach, which allows a higher probability of observing zeros compared to what is expected under the Poisson or Negative Binomial distributions. The data distribution can be expanded by incorporating either a Zero-Inflated Poisson (ZIP) or a Zero-Inflated Negative Binomial. In this study, we specifically consider the ZIP distribution [4], which combines two generating processes. The first process is a binomial with a probability p_g. If this binomial outcome is 0, the observed count is guaranteed to be zero with probability one. Otherwise, the counting follows a Poisson distribution. Therefore, we have the following probability mass function for $Y_{\cdot,cg} \sim ZIP(\mu_{\cdot,cg}, p_g)$:

$$\begin{aligned}
\mathbb{P}(Y_{\cdot,cg} = y) &= p_g \frac{\mu_{\cdot,cg}^y e^{-\mu_{\cdot,cg}}}{y!} \quad \text{for } y = 1, 2, \dots \\
\mathbb{P}(Y_{\cdot,cg} = 0) &= (1 - p_g) + p_g e^{-\mu_{\cdot,cg}}
\end{aligned}$$

2.1 Gene latent structure

RNA velocity modeling relies on a gene-specific Chemical Reaction Network (CRN) that delineates three processes: i) transcription of DNA into unspliced messenger RNA (mRNA); ii) splicing from unspliced mRNA to spliced mRNA; iii) degradation of spliced mRNA. Specifically, we are interested in the time-dynamic of spliced and unspliced mRNA, indicated as $s_g(\tilde{t}_c)$ and $u_g(\tilde{t}_c)$, where $\tilde{t}_c$ is the elapsed time since the formation of the cell c. We assume that the rates ruling the splicing and degradation, respectively $\beta_g \in \mathbb{R}^+$ and $\gamma_g \in \mathbb{R}^+$, remain constant over time, while the transcription rate is described using a piece-wise constant function

$$\alpha_g(\tilde{t}_c) = \begin{cases} \alpha_g^{\text{off}} & \text{if } 0 \leq \tilde{t}_c \leq t_{0,cg}^{\text{on}}; \\ \alpha_g^{\text{on}} & \text{if } t_{0,cg}^{\text{on}} \leq \tilde{t}_c \leq t_{0,cg}^{\text{on}} + t_{0,cg}^{\text{off}}; \\ \alpha_g^{\text{off}} & \text{if } \tilde{t}_c \geq t_{0,cg}^{\text{on}} + t_{0,cg}^{\text{off}}. \end{cases}$$

We emphasize that the piece-wise nature of the transcription rate captures the biological process governing cellular activity. A gene can undergo a repressive phase characterized by an absent or low transcription rate (α_g^{off}), resulting in a diminished production of associated proteins. Conversely, it can transition to an active phase represented by a constant higher rate (α_g^{on}), with $\alpha_g^{\text{on}} > \alpha_g^{\text{off}}$. The moment in which the transition of a gene from the repressive to the active phase occurs is denoted by $t_{0,cg}^{\text{on}}$. Subsequently, the gene remains activated for a duration of $t_{0,cg}^{\text{off}}$ time, until some biological conditions induce its return to the repressive phase. This specific moment $t_{0,cg}^{\text{off}}$, in which the modification is triggered, is called *switching time*.

For a fixed gene g, the time-dynamic is associated with the following ODE

$$\begin{cases} \frac{du_g(\tilde{t}_c)}{dt} = \alpha_g(\tilde{t}_c) - \beta_g u_g(\tilde{t}_c); \\ \frac{ds_g(\tilde{t}_c)}{dt} = \beta_g u_g(\tilde{t}_c) - \gamma_g s_g(\tilde{t}_c); \end{cases} \tag{1}$$

with initial conditions

$$s_g(0) = \frac{\alpha_g(0)}{\gamma_g} = \frac{\alpha_g^{\text{off}}}{\gamma_g} \qquad u_g(0) = \frac{\alpha_g(0)}{\beta_g} = \frac{\alpha_g^{\text{off}}}{\beta_g}.$$

The solution $(s_g(\tilde{t}_c), u_g(\tilde{t}_c))$ of the ODE system (1) is known in a closed form and is employed in the modelling phase.

2.2 Likelihood parameters and identifiability issue

To fully define the model, we need to specify how the parameter $\mu_{.,cg}$ is related to the ODE solution. The most natural assumption is to set $\mu_{s,cg} = s_g(\tilde{t}_c)$ and $\mu_{u,cg} = u_g(\tilde{t}_c)$, but this overlooks potential variations in the ability to measure mRNA across different cells. Therefore, we introduce a cell-specific parameter λ_c to account for distinct capture efficiencies. Consequently, we have

$$\mu_{s,cg} = \lambda_c s_g(\tilde{t}_c), \qquad \mu_{u,cg} = \lambda_c u_g(\tilde{t}_c)$$

It's important to note that the parameters λ_c cannot be identified simultaneously with the ODE parameters. This is due to the existence of multiple solutions that yield the same expected values $\mu_{.,cg}$. To address this, we assume $\lambda_1 = 1$. Similarly, the set of parameters $\left(\alpha_g^{\text{off}}, \alpha_g^{\text{on}}, \gamma_g, \beta_g, t_{0,cg}^{\text{on}}, t_{0,cg}^{\text{off}}, \tilde{t}_c\right)'$ cannot be jointly identified, and to resolve this, we set $\beta_g = 1$ and $t_{0,cg}^{\text{on}} = 0$.

An additional issue arises from the fact that, even though the cell-time can theoretically be estimated, single-cell RNA-seq data lacks sufficient information for such estimation. Therefore, we group cells using the cell-type labels, assuming a shared time for each member of the same cell-type. This approach is taken due to the destructive nature of the sequencing technique and to the inherit limitations of the data, as previously discussed [5].

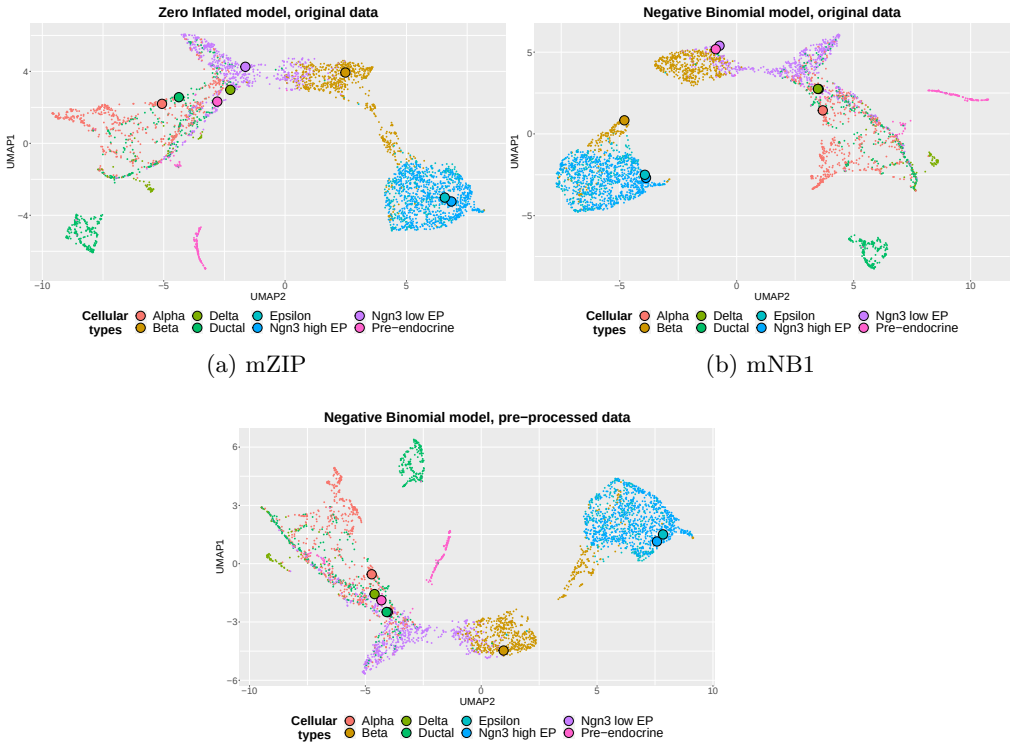
3 Results

To validate our approach, we aim to compare the results obtained using the Zero-Inflated Poisson and Negative Binomial assumption, without any preprocessing, and the ones derived with the Negative Binomial distribution under the standard preprocessing pipeline. The first two models are indicated as mZIP and mNB1, while the third one is mNB2.

We utilize the "pancreatic" dataset from the scVelo paper [1] and apply their standard filtering procedure. This procedure removes genes for which the sum of unspliced and spliced counts in all cells is smaller than 20. Following this, the selection is narrowed down to include only the top 2000 genes with the highest dispersion index. It is important to note that this initial preprocessing significantly reduces the frequency of zero observations. Due to computation-time constraints, for the NB-based model, we randomly select 1000 genes from the initial 2000 selected. Meanwhile, for our proposed model based on the ZIP distribution, we use the same 1000 genes and additionally include 500 genes from the list of those discarded by the preprocessing. The discarded genes mainly consist of zero counts. This strategy aims to assess the performance of our proposed model on genes that were initially filtered out due to the standard preprocessing, offering a comprehensive evaluation.

We estimated the model under a Bayesian framework with 250000 iterations, thinning at 25, and a burn-in of 20000 iterations, resulting in 2000 iterations for posterior computation. We define prior distributions on the coordinates of the steady states, since they offer greater interpretability in relation to gene dynamics. We assume $\frac{\alpha_g^{\text{on}}}{\beta_g}, \frac{\alpha_g^{\text{on}}}{\gamma_g} \sim$Beta$(2,1)$ and $\frac{\alpha_g^{\text{off}}}{\beta_g} \sim$Beta$(1,1)$. Regarding the switching times, we have $t_{0,cg}^{\text{off}} \sim$Exp$(1)$, while the time distribution is induced by the prior on the correspondent u-coordinate, i.e. we assume $u_g(\tilde{t}_c) \sim \text{Unif}\left(\frac{\alpha_g^{\text{off}}}{\beta_g}, u_{0,cg}^{\text{off}}\right)$, with $u_{0,cg}^{\text{off}} := u_g(t_{0,cg}^{\text{off}})$. Moreover, $p_g \sim$ Unif$(0,1)$, $\lambda_c \sim$ Unif$(0,1)$, $\eta_g \sim \mathcal{N}_{[0,+\infty)}(0, 10000)$, where $\mathcal{N}_{[a,b)}$ is a truncated Normal distribution with support in $[a,b)$.

As the data reside in a high-dimensional space, we project them onto a lower-dimensional space using UMAP procedures [7] for visualization purposes. Since



(a) mZIP (b) mNB1

(c) mNB2

Fig. 1: UMAP projections illustrating the scaled observations $(y_{s,cg}, y_{u,cg})/\lambda_c$ (small dots) alongside the estimated coordinates $(s_g(.), u_g(.))$ for each cell type (large dots). Distinct cellular types are distinguished by different colors. The dots are computed using the maximum-at-posterior values of λ_c, $s_g(.)$, and $u_g(.)$.

cell types are characterized by different gene expressions, we expect that the posterior means of $\mu_{s,cg}$ and $\mu_{u,cg}$, when projected onto the same UMAP plane, should exhibit clear separation.

Figure 1 presents the UMAP visualization, with panels (a) and (c) demonstrating comparable results in terms of cell-type separation. Notably, Figure 1 (b), corresponding to mNB1, fails to differentiate cell types adequately. This suggests that our proposed approach, which avoids arbitrary preprocessing, yields satisfactory results. Furthermore, the dots in Figure 1 (a) are more representative of their associated groups, as they tend to be located closer to the central position of the observations.

4 Conclusion

In this study, we have introduced preliminary findings regarding the application of zero-inflated distributions in the analysis of RNA-velocity models, eliminating the need for artificial preprocessing. These results were compared with other commonly employed likelihood models using a real dataset. Our findings indicate that, with our proposed approach, we were able to achieve results comparable to those obtained by a model based on the Negative Binomial distribution with a preprocessing stage. Currently, we are delving deeper into the investigation of the impact of our approach on the estimated parameters and velocity.

References

1. Bergen, V., Lange, M., Peidli, S., Wolf, F. A., Theis, F. J. (2020). Generalizing RNA velocity to transient cell states through dynamical modeling. Nature biotechnology, 38(12), 1408-1414.
2. Bergen, V., Soldatov, R. A., Kharchenko, P. V., Theis, F. J. (2021). RNA velocity-current challenges and future perspectives. Molecular systems biology, 17(8), e10282.
3. Gorin, G., Fang, M., Chari, T., Pachter, L. (2022). RNA velocity unraveled. PLOS Computational Biology, 18(9), e1010492.
4. Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics, 34(1), 1-14.
5. Sabbioni, E., Bibbona, E., Mastrantonio, G., Guido, S. (2023). Analyzing RNA data with scVelo: identifiability issues and a Bayesian implementation. In Book of the Short Papers SEAS IN 2023 (pp. 538-543). PEARSON.
6. Jahnke, T., Huisinga, W. (2007). Solving the chemical master equation for monomolecular reaction systems analytically. Journal of mathematical biology, 54, 1-26.
7. McInnes et al., (2018). UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software, 3(29), 861, https://doi.org/10.21105/joss.00861