Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling

ZHANG,Zihao

December 15, 2024

1 Introduction

The paper *Generalizing RNA Velocity to Transient Cell States through Dynamical Modeling* is particularly interesting to me because it combines the Expectation-Maximization (EM) algorithm with partial differential equations (PDEs) to address the biological challenge of estimating RNA velocity in dynamic cellular states. RNA velocity is a measure of the rate of change in gene expression over time, which provides valuable insights into cellular dynamics and differentiation processes. However, the conventional steady-state models for RNA velocity have limitations in capturing the transient and dynamic nature of gene expression in cells.

The proposed method in this paper introduces a dynamic model to address these limitations. By incorporating latent variables for time and cell state, the method allows for more accurate modeling of gene expression changes during transitions between different cellular states. The dynamic model produces elegant and visually compelling results, showing how cells evolve over time and how their gene expression levels change accordingly.

2 Background and Problem Modeling

In order to study RNA velocity, it is essential to understand the system of differential equations that govern the transcription and splicing of mRNA. The process can be described as follows: the unspliced mRNA $u(t)$ is transcribed from the DNA by the transcription rate $\alpha^{(k)}$ in cell state k, and then it is spliced into the mature, spliced mRNA $s(t)$ at a rate β. Finally, both $u(t)$ and $s(t)$ undergo degradation, with degradation rates γ for the spliced and unspliced mRNA, respectively. This process can be expressed using the following reactions:

$$\emptyset \xrightarrow{\alpha^{(k)}} u(t) \xrightarrow{\beta} s(t) \xrightarrow{\gamma} \emptyset$$

In this system, $u(t)$ represents the unspliced mRNA, $s(t)$ represents the spliced mRNA, and the arrows indicate the transitions between different states of mRNA due to transcription, splicing, and degradation. The rate constants $\alpha^{(k)}$, β, and γ govern these processes. $\alpha^{(k)}$ depends on the cell state k since different cell states might have different transcription rates, while β and γ are assumed to be constant.

The standard steady-state models assume that these variables evolve toward a stable equilibrium. However, this assumption fails when the system is in a transient state or undergoing rapid changes, such as during differentiation or in response to external stimuli. In these dynamic cases, the steady-state assumption does not adequately describe the cellular behavior, as gene expression levels do not stabilize but rather continue to evolve over time.

To address this limitation, the paper proposes a dynamic model that tracks cells through different states over time, allowing for a more accurate representation of RNA dynamics during transitions between cellular states. This model is based on a system of differential equations that capture the continuous changes in the unspliced and spliced mRNA over time.

The key insight is that RNA splicing is a time-dependent process that occurs continuously as the cell transitions through different states. By modeling this process dynamically, we can capture the varying rates of transcription and splicing at different points in time, providing a more flexible and accurate framework for estimating RNA velocity. This dynamic approach allows us to model how cells move from one state to another over time, as opposed to assuming a constant steady-state equilibrium. The dynamics of RNA splicing over time can thus effectively track the transitions in gene expression as cells undergo differentiation, responses to stimuli, or other time-dependent processes.

3 Challenges in Steady-State Models

Steady-state models, such as velocyto, make several simplifying assumptions that hinder their ability to capture transient cellular dynamics. These assumptions include:

- **Constant transcription and splicing rates:** These models assume that the transcription and splicing rates are constant over time, which is often not true in biological systems where these rates can change in response to external factors.
- **Absence of cell state transitions:** Steady-state models do not account for the fact that cells can transition between different states (e.g., from active to inactive or from one differentiation stage to another).
- **Simplified temporal dynamics:** The time-dependent nature of RNA dynamics is often neglected, limiting the ability to capture transient cellular behaviors.

In the steady-state model (velocyto), RNA velocities are estimated as the deviation from a steady-state model fit. Specifically, velocities are computed by fitting a linear regression model to extreme quantiles of cell data (cells representing steady states). The velocities v_i are then calculated as the difference between the unspliced mRNA u_i and the steady-state model's prediction $\gamma' s_i$ for the spliced mRNA s_i:

$$u_\infty \approx \gamma' s_\infty \quad (\beta = 1)$$
$$v_i = u_i - \gamma' s_i$$

However, this approach has several potential drawbacks:

- **Assumes sampled steady states:** The model assumes that the data represent sampled steady states, which may not be the case in dynamic processes.
- **Not all samples are used:** Only extreme quantile cells (representing steady states) are considered, potentially excluding important transient data points.
- **Not fully identifiable:** The model is not fully identifiable because it cannot separately infer the relative transcription rate and common splicing rate.

These challenges highlight the limitations of steady-state models in capturing the transient and complex dynamics of RNA velocity. This motivates the need for a more flexible dynamic model that accounts for the continuous transition between cellular states over time.

4 Dynamic Model and Solution

4.1 Deterministic Model

This section details the deterministic model for gene expression dynamics. Based on the law of mass action, we construct the following differential equation system:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = \alpha(t) - \beta u(t) \ \frac{\mathrm{d}s}{\mathrm{d}t} \ = \beta u(t) - \gamma s(t) \tag{1}$$

Where $\alpha(t)$ is defined as:

$$\alpha(t) = \begin{cases} \alpha^{\mathrm{on}}, & t \leq t_s \ \alpha^{\mathrm{off}} = 0, \\ t > t_s \end{cases} \tag{2}$$

RNA velocity $\boldsymbol{v}(t)$ is defined as:

$$\boldsymbol{v}(t) = \left(\frac{\mathrm{d}s_g}{\mathrm{d}t}\right)_g = \left(\beta_g u_g(t) - \gamma_g s_g(t)\right)_g \in \mathbb{R}^{n_g} \tag{3}$$

4.2 Scale Invariance

The model exhibits scale invariance property, such that for any scaling parameter $\kappa > 0$:

$$\left(u\left(t; \theta_r, t_s\right), s\left(t; \theta_r, t_s\right)\right) = \left(u\left(\kappa t; \theta_r/\kappa, \kappa t_s\right), s\left(\kappa t; \theta_r/\kappa, \kappa t_s\right)\right) \tag{4}$$

Scale invariance implies parameter inference degeneracy. To ensure well-posedness of the inference, we need to fix the system's time scale, for example, by studying the dynamic process within a fixed period $[0, T]$.

5 Expectation-Maximization (EM) Algorithm

5.1 EM Algorithm Framework

Given the observed data $X = (x_{cg})_{c=1:n_c;g=1:n_g}$, where $xcg = (u_{cg}, s_{cg})$, our objective is to maximize the log-likelihood function:

$$L(\theta; X) = \ln \prod_{c=1}^{n_c} \prod_{g=1}^{n_g} P(x_{cg} \mid \theta_g) = \sum_{c=1}^{n_c} \sum_{g=1}^{n_g} \ln P(x_{cg} \mid \theta_g) \tag{5}$$

By introducing the latent variable $h_{cg} = (t_{cg}, x_{s,cg})$, the log-likelihood function can be rewritten as:

$$l(\theta_g; x_{cg}) = \ln P(x_{cg} \mid \theta_g) = \ln P(x_{cg}, h_{cg} \mid \theta_g) - \ln P(h_{cg} \mid x_{cg}; \theta_g) \tag{6}$$

5.2 EM Algorithm Derivation

For the given probability model, assume the observation noise for each cell and gene follows a Gaussian distribution with variance σ^2, and sampling time t_{cg} is uniformly distributed over $[0, T]$:

$$P(x_{cg}, t_{cg} \mid \theta_g) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{|x_{cg} - x(t_{cg}; \theta_g)|^2}{2\sigma^2}\right) \cdot \frac{1}{T} \tag{7}$$

E-step calculates the conditional expectation of latent variables:

$$Q(\theta \mid \theta^{(j)}) = \mathbb{E}_{h|X,\theta^{(j)}}(L_0(\theta; X, h)) \tag{8}$$

M-step updates parameters by maximizing $Q(\theta \mid \theta^{(j)})$:

$$\theta^{(j+1)} = \arg\min_{\theta} \int_0^T |X - X(t; \theta)|^2 \exp\left(-\frac{|X - X(t; \theta^{(j)})|^2}{2\sigma^2}\right) dt \tag{9}$$

In the small noise limit, this can be simplified to:

$$\theta^{(j+1)} = \arg\min_{\theta} |X - X(t^{(j)}; \theta)|^2 \tag{10}$$

Where $t^{(j)}$ is estimated by:

$$t^{(j)} = \arg\min_{t} |X - X(t; \theta^{(j)})|^2 \tag{11}$$

6 Simulation Methodology for EM Algorithm Validation

6.1 Data Generation Strategy

To validate the EM algorithm for RNA velocity estimation, we simulated synthetic gene expression data following realistic transcription dynamics with noise injection.

6.1.1 Simulation Setup

- **Number of cells:** $n = 600$
- **Number of genes:** $d = 15$
- **Transcription stages:** On-stage (active transcription) and Off-stage (transcription turned off)
- **Noise model:** Gaussian noise $\mathcal{N}(0, 0.25)$

6.1.2 Gene Expression Dynamics

The dynamics of the unspliced RNA (u) and spliced RNA (s) were generated as follows:

On-Stage Dynamics. During active transcription, we used:

$$\begin{aligned} u(t) &= u_0 e^{-\beta t} + \frac{\alpha}{\beta}(1 - e^{-\beta t}), \\ s(t) &= s_0 e^{-\beta t} + \frac{\alpha}{\beta}(1 - e^{-\beta t}) - (\alpha - \beta u_0) t e^{-\beta t}. \end{aligned} \tag{12}$$

- $\alpha \sim \text{Uniform}[15, 35]$ (transcription rate).
- $\beta = 1.2$ (splicing rate).
- $u_0 = 0, s_0 = 0$ (initial conditions).

Off-Stage Dynamics. After a switch time t_{switch}, transcription stops, and the dynamics follow:

$$\begin{aligned} u(t) &= u_{\text{switch}} e^{-\beta(t - t_{\text{switch}})}, \\ s(t) &= s_{\text{switch}} e^{-\gamma(t - t_{\text{switch}})} - \frac{\beta u_{\text{switch}}}{\gamma - \beta} \left(e^{-\gamma(t - t_{\text{switch}})} - e^{-\beta(t - t_{\text{switch}})} \right), \end{aligned} \tag{13}$$

- $\gamma \sim \text{Uniform}[1.0, 2.5]$ (degradation rate).
- $t_{\text{switch}} \sim \text{Uniform}[0.5, 1.5]$ (transition between stages).

Sampling and Noise.

- Time points t were sampled uniformly from $[0, T]$, where $T = 2\ln(12)$ ensures no steady-state stabilization.
- Gaussian noise $\mathcal{N}(0, 0.25)$ was added to simulate realistic observational noise.

6.2 Evaluation and Results

To evaluate the EM algorithm's performance, the following procedure was performed:

1. **Parameter Inference:** Using the EM algorithm, the kinetic parameters (α, γ) and latent variables were inferred.
2. **RNA Velocity Estimation:** The estimated RNA velocities $(\hat{u}, \hat{s})$ were compared against the true synthetic values (u, s).
3. **Visualization of Results:** The accuracy of the EM algorithm was visualized using scatter plots of the true versus estimated RNA velocities. To validate the fitting performance:
    - True values are plotted as blue scatter points.
    - Estimated values are plotted as red scatter points.
    - A dashed reference line represents perfect agreement.

The comparison is shown in Figure 1, where the alignment of the estimated values (red) with the true values (blue) demonstrates the algorithm's effectiveness.

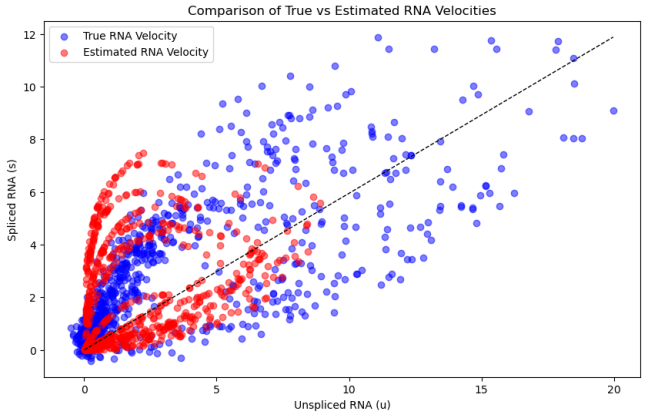

Figure 1: Comparison of true RNA velocities (blue) and estimated RNA velocities (red). The values cluster around the dashed reference line, indicating good agreement between true and estimated velocities.

6.3 Analysis of Results

The results in Figure 1 demonstrate that the EM algorithm successfully recovers RNA velocities under the simulated dynamics. Key observations include:

- The red scatter points (estimated values) align closely with the blue scatter points (true values), validating the accuracy of the algorithm.
- Both the unspliced and spliced RNA values exhibit minimal deviation from the dashed reference line, indicating that the kinetic parameters and latent variables were well inferred.

- Minor deviations may result from observational noise or model assumptions, but overall, the EM algorithm performs robustly across cells and genes.

7 Conclusion

This report introduces the dynamic RNA velocity model proposed by V. Bergen [1]. However, the probabilistic framework for the dynamic model was not explicitly detailed in their work. To address this, I adopted the mathematical framework established by T. Li [2], providing a thorough analysis of the EM algorithm within a deterministic setting. Simulations were conducted to evaluate the effectiveness of the algorithm. The code used for this study was implemented by myself and is available at: `https://github.com/arcsin231/MATH5472-final-project-ZZH`.

References

[1] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, "Generalizing rna velocity to transient cell states through dynamical modeling," *Nature biotechnology*, vol. 38, no. 12, pp. 1408–1414, 2020.

[2] T. Li, J. Shi, Y. Wu, and P. Zhou, "On the mathematics of rna velocity i: theoretical analysis," *bioRxiv*, pp. 2020–09, 2020.