CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2023-0026

Vol. **5**, No. 1, pp. 182-220
March 2024

On the Mathematics of RNA Velocity II: Algorithmic Aspects

Tiejun Li[1,2,3,*], Yizhuo Wang[3], Guoguo Yang[1] and Peijie Zhou[2,4]

[1] *LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, China.*
[2] *Center for Machine Learning Research, Peking University, Beijing 100871, China.*
[3] *Center for Data Science, Peking University, Beijing 100871, China.*
[4] *Department of Mathematics, University of California, Irvine, CA 92697, USA.*

Received 5 June 2023; Accepted 5 December 2023

Abstract. In the previous paper [CSIAM Trans. Appl. Math. 2 (2021), 1–55], the authors proposed a theoretical framework for the analysis of RNA velocity, which is a promising concept in scRNA-seq data analysis to reveal the cell state-transition dynamical processes underlying snapshot data. The current paper is devoted to the algorithmic study of some key components in RNA velocity workflow. Four important points are addressed in this paper: (1) We construct a rational time-scale fixation method which can determine the global gene-shared latent time for cells. (2) We present an uncertainty quantification strategy for the inferred parameters obtained through the EM algorithm. (3) We establish the optimal criterion for the choice of velocity kernel bandwidth with respect to the sample size in the downstream analysis and discuss its implications. (4) We propose a temporal distance estimation approach between two cell clusters along the cellular development path. Some illustrative numerical tests are also carried out to verify our analysis. These results are intended to provide tools and insights in further development of RNA velocity type methods in the future.

AMS subject classifications: 92B05, 92-08, 92-10

Key words: Time-scale fixation, uncertainty quantification, optimal kernel bandwidth, temporal distance estimation.

1 Introduction

The development of single-cell RNA sequencing (scRNA-seq) technology has revolutionized the resolution and capability to dissect the cell-fate determination process [42]. How-

*Corresponding author. *Email addresses:* ygj512@hotmail.com (G. Yang), jiuqie@pku.edu.cn (Y. Wang), tieli@pku.edu.cn (T. Li), peijiez1@uci.edu (P. Zhou)

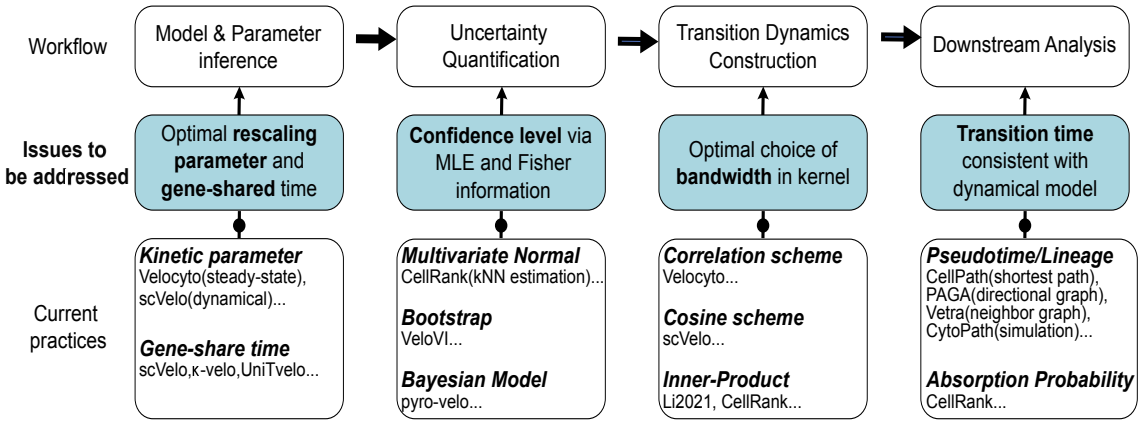

Figure 1: The computational workflow of RNA velocity analysis and under-addressed issues.

ever, traditional scRNA-seq datasets only provide static snapshots of gene expression among cells at a certain time point, which lack the direct temporal information to infer the dynamics of cell state transitions [47]. To address this limitation, the RNA velocity method [21] utilizes both unspliced and spliced counts in scRNA-seq data to model and infer the dynamics of mRNA expression and splicing process, allowing the prediction of gene expression changes over time, and the specification of directionality during development. The method has been applied widely in different biological systems [1, 9, 12], and the computational workflow of RNA velocity analysis has been established and undergone rapid development [3, 21, 23, 50] (Fig. 1).

To improve the effectiveness and robustness of RNA velocity analysis, various algorithmic modifications have been proposed throughout the computational workflow. For the parameter inference step, scVelo utilizes an Expectation-Maximization (EM) procedure between latent time specification and kinetic parameter update to generalize the steady-state assumption to the transient dynamical process [3]. In addition, κ-velo proposes to calculate a gene-shared latent time for each cell by approximating the traveling time with the number of cells in-between [29], and UniTvelo calculates the unified latent time by aggregating the gene-specific time quantiles [10]. Recently, VeloVAE utilizes variational Bayesian inference and autoencoder to compute the gene-shared latent time and cell latent state [15]. To account for the uncertainty of inferred parameters incurred by noise and sparsity in spliced or unspliced counts, CellRank adopts the multivariate normal model to quantify the velocity distribution [23], while VeloVI employs the bootstrap strategy [11]. Recently, pyro-Velo proposes a Bayesian approach to model the posterior distribution of parameters [35].

Based on the inferred RNA velocity, downstream dynamical analysis tools such as low-dimensional embedding [1, 34] and trajectory inference [10, 27, 50] are developed by leveraging the cell-cell neighbor graph directed by the velocities. Pertinent to such methods is the construction of a cellular random walk transition probability (or weight)