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)

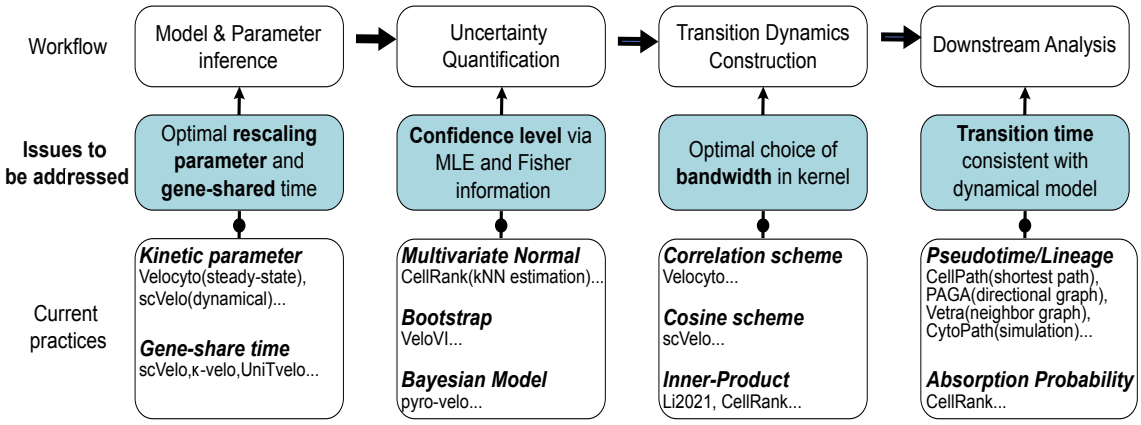

Figure 1: The computational workflow of RNA velocity analysis and under-addressed issues.

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)

matrix, which is induced by the velocity-based data kernels. Our previous theoretical work [26] has elucidated that different choices of velocity kernels can result in various forms of differential equations as the continuum limit. The resulting Markov chain of cell-state transitions can provide further insights into the underlying dynamics, including the transition paths to quantify development routes [26,52], and the absorption probabilities to quantify cell-fate commitment likelihood [23,37].

Despite the success of algorithm developments, deeper analysis is still necessary to understand the rationales of the algorithm design and address the unresolved issues. For instance, in parameter estimation, the optimal choice of re-scaling parameter and therefore the determination of gene-shared latent time beyond heuristic strategies have yet to be determined. In fact, extracting the hidden time information from the snapshot data is still a central issue in scRNA-seq data analysis. Meanwhile, the uncertainty quantification of parameters could benefit from a rigorous confidence level analysis of the EM algorithm. In the construction of random walks, the appropriate choice of kernel bandwidth relies on the numerical analysis of convergence order to continuum limit. In addition, simulations and benchmarks are important to validate the statistical and numerical analysis results.

As a continuation of our previous work [26], in this paper we will study the mathematics of RNA velocity analysis from the perspective of algorithms. The main contributions of this paper toward the current computation workflow could be summarized as follows:

- **Parameter Inference.** To determine the gene-latent time [3], we formulate an optimization framework to determine the gene-specific rescaling parameters and propose the numerical scheme to efficiently tackle the considered problem.

- **Uncertainty Quantification.** Employing the asymptotic theory for EM algorithm [32], we rigorously derive the confidence level for the kinetic parameters in the dynamical RNA velocity model.

- **Random Walk Construction.** Addressing the finite sample-size issue in computation, we analyze the variance and bias of the approximation to continuum equation, and find an optimal criterion to determine kernel bandwidth and sample size for the velocity kernel which induces the cellular random walk dynamics.

- **Downstream Analysis.** To perform lineage inference that is consistent with RNA velocity dynamics, we propose the first hitting time analysis to quantify the transition time.

The rest of this paper is organized as follows. According to the contents stated above, we have studied them in Sections 2, 3, 4 and 5, respectively, and given corresponding numerical illustrations in each section. Finally, we make the conclusion. Some proof details are left in the Appendices A and B.

2 Fixation of time rescaling constants

The current RNA velocity models [3, 21] assume that the genes are independent, thus in dynamical parameter inference [3], the latent times of the cells are obtained for each gene independently, and we lack a determination step of time-rescaling factors to form a globally consistent gene-shared time. To rationally infer the global cell hidden times as well as determine the rescaling factors, we propose an optimization framework to address this issue. It could be served as a good starting point for more delicate inference on the latent time by taking into account more technical details.

2.1 Problem setup

Suppose that in a considered scRNA-seq measurement, we have d genes with the label $g=1,2,\ldots,d$ and n cells with the label $c=1,2,\ldots,n$. Similar to the previous work, we utilize the deterministic dynamical model

$$\begin{aligned}\frac{\mathrm{d}u}{\mathrm{d}t}&=\alpha^{\mathrm{on/off}}(t)-\beta u(t),\\ \frac{\mathrm{d}s}{\mathrm{d}t}&=\beta u(t)-\gamma s(t)\end{aligned}\tag{2.1}$$

to describe the transcriptional process of each gene, and the individual genes are independent of each other. Here $t\geq 0,(u(t),s(t))|_{t=0}=(u_0,s_0)$, and

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

where t_s is the switching time of the transcriptional process at which transcription rate α turns to 0. The variables $u(t)$ and $s(t)$ are the abundance of unspliced and spliced mRNA in the cell measured at time t, respectively. In general, the resulting data are not time-resolved and t is a latent variable. Likewise, the transcriptional state of the cell (on/off) is an unknown variable, and the rates $\alpha^{\mathrm{on}},\beta$ and γ cannot be directly measured experimentally.

In the inference process, we need to solve the equation and infer the kinetics of splicing controlled by parameters: transcription rate α^{on}, splicing rate β and degradation rate γ, latent variable time t. We usually infer the parameters for each gene separately under the independent gene assumption, which leaves the relative size of the parameters of genes as an unsolved problem. As the system has the following scale invariance property [26], i.e. if we define the parameter $\theta=(\theta_r,t_s)$, in which $\theta_r=(\alpha,\beta,\gamma)$, then the following equation holds:

$$\big(u(t;\theta_r,t_s),s(t;\theta_r,t_s)\big)=\big(u(\kappa t;\theta_r/\kappa,\kappa t_s),s(\kappa t;\theta_r/\kappa,\kappa t_s)\big),\tag{2.2}$$

where $\kappa>0$ is the scaling parameter. In the inference we usually keep $\beta_g=1$ at first while optimizing other parameters for each specific gene g, which essentially infers α_g/β_g

and γ_g/β_g due to the scale invariance. When considering the high-dimensional velocity and the corresponding low-dimensional projection, the scale needs to be adjusted among all genes, that is, the scaling parameters $(\beta_g)_g$ need to be determined for each gene. As this parameter appears in the final RNA velocity, its choice will highly affect the lineage inference in downstream analysis. Also, computing the gene-latent time required us to find out the scaling parameters. Indeed, this is an important under-addressed issue in scRNA-seq data analysis.

Assume that we have already inferred the unscaled parameters $\alpha_g, \beta_g = 1, \gamma_g$ for each gene, with the gene-specific cell time matrix $T = (t_{cg}) \in (\mathbb{R}^+ \cup \{0\})^{n\times d}$. Our goal is to infer the gene-shared latent time t_c for each cell as well as determine the rescaling parameters β_g for different genes. Below we will propose two optimization approaches to tackle this issue.

2.2 Gene-shared time through optimization

To obtain the gene-shared latent time in any given cell, we reason it to be as consistent as possible with the respective rescaled time for each gene within the cell. Denote by $\beta = (\beta_g)$ or $x = (x_g) = (\beta_g^{-1}) \in \mathbb{R}^d$ the time re-scaling parameters for the genes, and $t = (t_c) \in \mathbb{R}^n$ the gene-shared latent time for cells to be optimized. We formulate the above consistency intuition through two proposals.

Our first proposal is based on the model

$$t_{cg}\beta_g^{-1} = t_c + \epsilon_{cg}, \quad \epsilon_{cg} \sim N(0,\sigma^2), \quad c = 1,\dots,n, \quad g = 1,\dots,d. \tag{2.3}$$

Here t_{cg} is the inferred gene-specific time with $\beta_g = 1$, and $t_{cg}\beta_g^{-1}$ is the rescaled time with β_g, and (2.3) means that the rescaled time should be consistent with a global gene-shared common time t_c upon removing some noise. With this setup, we can determine x and t with the following formulation.

Proposal 2.1 (Inference with Multiplicative Noise)**.** *The gene-shared latent time t and re-scaling parameters x can be determined through the minimization problem*

$$(x^*, t^*) = \underset{\|t\|=1; x\succ 0, t\succeq 0}{\operatorname{argmin}} \|TX - t\mathbf{1}^\top\|_F^2, \tag{2.4}$$

where $x \succ 0, t \succeq 0$ means that x, t have positive or non-negative components, respectively, $\mathbf{1} = (1,1,\dots,1)^\top \in \mathbb{R}^d$, $X = \operatorname{diag}(x_1,\dots,x_d) \in \mathbb{R}^{d\times d}$ is the diagonal matrix formed by the components of x, and

$$\|A\|_F := \left(\sum_{ij} a_{ij}^2\right)^{\frac{1}{2}} = \left(\operatorname{tr}(AA^\top)\right)^{\frac{1}{2}}, \quad A = (a_{ij})$$

denotes the Frobenius norm (F-norm) of a matrix.

Theorem 2.1. *Assume that the inferred gene-specific cell time matrix T satisfies the condition*

$$T\in\mathcal{T}=\left\{T\in\left(\mathbb{R}^{+}\cup\{0\}\right)^{n\times d}\,\middle|\,T^{\top}T \text{ is irreducible}\right\}. \tag{2.5}$$

Then the optimization problem (2.4) *has the unique solution $x^{*}=d\lambda_1^{-1/2}Wv_1$, where v_1 is the ℓ^2-unit eigenvector corresponding to the maximal eigenvalue λ_1 of*

$$H=W^{\top}T^{\top}TW,\quad W:=\mathrm{diag}(w_1,\ldots,w_d),\quad w_g=\frac{1}{\|t_{\bullet g}\|},\quad g=1,\ldots,d, \tag{2.6}$$

and it has positive components. The global gene-shared common time

$$t^{*}=\frac{TWv_1}{\|TWv_1\|}.$$

Proof. To solve the problem (2.4), we note that when x is fixed, the optimization

$$\min_{t\succeq 0}\|TX-t\mathbf{1}^{\top}\|_F^2$$

turns out to be a least squares problem, and the minimum point is $t=Tx/d$. Substituting it back to (2.4), we get

$$x^{*}=\underset{\|Tx\|=d,x\succ 0}{\mathrm{argmin}}\left\|TX-\frac{1}{d}TXE\right\|_F^2,$$

where the matrix $E:=\mathbf{1}\mathbf{1}^{\top}\in\mathbb{R}^{d\times d}$. Define $C=I_d-E$, which satisfies $C^{\top}=C$ and $C^2=C$. We have

$$x^{*}=\underset{\|Tx\|=d,x\succ 0}{\mathrm{argmin}}\ \|TXC\|_F^2.$$

By the definition of the F-norm, we have

$$x^{*}=\underset{\|Tx\|=d,x\succ 0}{\mathrm{argmin}}\ \mathrm{tr}(TXCC^{\top}X^{\top}T^{\top}). \tag{2.7}$$

Denote by $A\circ B$ the Hadamard product of matrices A and B defined as $A\circ B=(a_{ij}b_{ij})$ for $A=(a_{ij})$ and $B=(b_{ij})$. It is not difficult to find that (2.7) is equivalent to

$$x^{*}=\underset{\|Tx\|=d,x\succ 0}{\mathrm{argmin}}\ x^{\top}Mx, \tag{2.8}$$

where $M=(T^{\top}T)\circ C$.

Denote by $t_{\bullet g}$ the vector formed by $(t_{cg})_c$ for a fixed gene g. By the irreducibility condition (2.5), we have $\|t_{\bullet g}\|>0$ for any g, thus W is well-defined. Further note that

$$M=(T^{\top}T)\circ C=W^{-2}-\frac{1}{d}T^{\top}T,$$

the problem (2.8) is equivalent to

$$x^* = \underset{\|Tx\|=d, x\succ 0}{\operatorname{argmin}} \; x^\top W^{-2} x. \tag{2.9}$$

Suppose

$$H = Q^\top \Lambda Q, \quad Q^\top = (v_1, v_2, \ldots, v_d),$$

where $Q^\top Q = I_d$, $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_d)$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$. We have $Hv_k = \lambda_k v_k$ for $k = 1, \ldots, d$. Ignoring the positivity constraint $x \succ 0$, we can find that the optimizer of (2.9)

$$x^* = d\lambda_1^{-\frac{1}{2}} W v_1$$

by taking the transformation $z = QW^{-1}x$. Furthermore, the positivity of x can be guaranteed by the Perron-Frobenius theorem [19] for the non-negative and irreducible matrix H.

The optimal t^* is obtained by the relation

$$t^* = \frac{Tx^*}{d} = \lambda_1^{-\frac{1}{2}} TWv_1 = \frac{TWv_1}{\|TWv_1\|}.$$

The proof is done. □

Remark 2.1. The formulation (2.4) realizes the inference of model (2.3) through maximum likelihood estimation. The rescaling parameter x_g corresponds to the inverse splicing rate β_g^{-1}, and the normalization $\|t\| = 1$ is to fix the undetermined global time scale of the whole system. The constant $d\lambda_1^{-1/2}$ in x^* is not important but the orientation Wv_1 is essential.

Our second proposal is slightly different from the first one, and it is based on the model

$$t_{cg} = t_c \beta_g + \epsilon_{cg}, \quad \epsilon_{cg} \sim N(0, \sigma^2), \quad c = 1, \ldots, n, \quad g = 1, \ldots, d. \tag{2.10}$$

With this setup, we can directly determine β and t through the following maximum likelihood formulation.

Proposal 2.2 (Inference with Additive Noise)**.** *The gene-shared latent time t and the splicing rate β can be determined by solving the minimization problem*

$$(\beta^*, t^*) = \underset{\|\beta\|=1; \beta\succ 0, t\succeq 0}{\operatorname{argmin}} \; \|T - t\beta^\top\|_F^2. \tag{2.11}$$

Theorem 2.2. *Assume that the inferred gene-specific cell time matrix T satisfies the condition* (2.5)*. Then the optimization problem* (2.11) *has the unique solution $\beta^* = v_1$, where v_1 is the ℓ^2-unit eigenvector corresponding to the maximal eigenvalue λ_1 of*

$$H = T^\top T, \tag{2.12}$$

and it has positive components. The global gene-shared common time

$$t^* = Tv_1.$$

Proof. Note that when β is fixed, the optimization

$$\min_{t\in\mathbb{R}^d}\|T-t\beta^\top\|_F^2$$

is a least squares problem, and the minimum point is $t=T\beta/\|\beta\|^2$. Substituting it back and ignoring the normalization and positivity constraints on β at first, we obtain

$$\min_{\beta\in\mathbb{R}^d}\left\|T-\frac{T\beta\beta^\top}{\|\beta\|^2}\right\|_F^2 \iff \max_{\beta\in\mathbb{R}^d}\frac{\beta^\top T^\top T\beta}{\|\beta\|^2}.$$

The rates β can be determined up to a multiplicative constant. So we naturally take the normalization $\|\beta\|=1$ and consider the equivalent problem

$$\beta^*=\underset{\|\beta\|=1,\beta\succ 0}{\arg\max}\ \beta^\top T^\top T\beta. \tag{2.13}$$

By the condition (2.5) and the Perron-Frobenius theorem applied to the matrix $H=T^\top T$, the optimizer of (2.13) is unique and characterized by the unit eigenvector v_1 associated with the maximal eigenvalue λ_1 of H, and it has positive components. □

With the above proposals, we get the splicing rates β^* and gene-shared latent time t^*. We can make the rescaling

$$(\alpha_g,1,\gamma_g;t_{cg}) \longrightarrow (\alpha_g\beta_g^*,\beta_g^*,\gamma_g\beta_g^*;t_{cg}/\beta_g^*),\quad g=1,\dots,d$$

to get more reasonable parameters with the obtained β^*.

Remark 2.2. In this remark, we present the rationale of two proposed statistical models through some mathematical reasoning. Assume that the observation $x_{cg}=x(t_c;\theta_g)+\xi_{cg}$, where $x(t;\theta):=(u(t;\theta),s(t;\theta))$ is the solution of the deterministic mRNA expression dynamics (2.1) at time t with parameter θ, the noise $\xi_{cg}\sim\mathcal{N}(0,\sigma_x^2)$ and $\sigma_x\ll 1$. Due to the scale invariance, we know that $x(t_c;\theta_g)=x(t_c\beta_g;\tilde{\theta}_g)$ with $\tilde{\theta}_g:=(\alpha_g/\beta_g,1,\gamma_g/\beta_g;t_{s,g}\beta_g)$, which is also the working setup by setting $\beta_g=1$ at first in practical computations. The parameters of RNA velocity model are derived from EM algorithm [3, 26], which is also detailed in Section 3. From the E-step, we obtain

$$t_{cg}=\underset{t\geq 0}{\arg\min}\left\|x_{cg}-x(t;\hat{\tilde{\theta}}_g)\right\|^2, \tag{2.14}$$

where $\hat{\tilde{\theta}}_g$ is the final estimator of $\tilde{\theta}_g$ in EM iterations. From the large sample theory for the point estimation [24], we have

$$\hat{\tilde{\theta}}_g\approx\tilde{\theta}_g+\frac{1}{\sqrt{n}}\eta_g,\quad \eta_g\sim\mathcal{N}(0,\Sigma_g).$$

For the optimization problem (2.14), let $f(t)=\|x_{cg}-x(t;\hat{\tilde{\theta}}_g)\|^2$, then t_{cg} is the solution of the Euler-Lagrange equation

$$f'(t)=\Big(x_{cg}-x\big(t;\hat{\tilde{\theta}}_g\big)\Big)\cdot\frac{\mathrm{d}x\big(t;\hat{\tilde{\theta}}_g\big)}{\mathrm{d}t}=0.$$

Define

$$F\big(t;x_{cg},\hat{\tilde{\theta}}_g\big)=\big(x_{cg}-x\big(t;\hat{\tilde{\theta}}_g\big)\big)\cdot\frac{\mathrm{d}x\big(t;\hat{\tilde{\theta}}_g\big)}{\mathrm{d}t}$$

and note that

$$F\big(t_c\beta_g;x\big(t_c\beta_g;\tilde{\theta}_g\big),\tilde{\theta}_g\big)=0,$$

since $f(t)$ achieves the minimum 0 with such choice of parameters. Further assume that

$$\partial_t F\big(t_c\beta_g;x\big(t_c\beta_g;\tilde{\theta}_g\big),\tilde{\theta}_g\big)\neq 0,$$

then by the implicit function theorem, there exists a function $G(\cdot,\cdot)$ such that when (y_{cg},μ_g) belongs to a small neighborhood of $(x(t_c\beta_g;\tilde{\theta}_g),\tilde{\theta}_g)$, we have

$$t=G(y_{cg},\mu_g),\quad t_c\beta_g=G\big(x\big(t_c\beta_g;\tilde{\theta}_g\big),\tilde{\theta}_g\big). \tag{2.15}$$

Thus, we obtain

$$t_{cg}=G\big(x_{cg},\hat{\tilde{\theta}}_g\big)=G\left(x\big(t_c\beta_g;\tilde{\theta}_g\big)+\xi_{cg},\tilde{\theta}_g+\frac{1}{\sqrt{n}}\eta_g\right).$$

With Taylor expansion, we get

$$t_{cg}=G\big(x\big(t_c\beta_g;\tilde{\theta}_g\big),\tilde{\theta}_g\big)+\partial_x G\cdot\xi_{cg}+\partial_{\tilde{\theta}}G\cdot\frac{1}{\sqrt{n}}\eta_g+\text{h.o.t.},$$

which shows

$$t_{cg}=t_c\beta_g+\text{Gaussian noise}+\text{h.o.t.}$$

by (2.15) and the asymptotic independence between ξ_{cg} and η_g (note the Gaussianity of η_g weakly depends on one specific sample x_{cg}).

The above reasoning provides the rationale why we propose the two optimization formulations in this paper. Since $\partial_x G$ and $\partial_{\tilde{\theta}}G$ may depend on β_g, it is difficult to judge which one is more reasonable a priori. In addition, it should be noted that the above analysis is based on the assumption that the observation noise is Gaussian. If the noise is non-Gaussian and only has zero mean, our derivation still holds, but

$$t_{cg}=t_c\beta_g+\text{zero mean noise}+\text{h.o.t.}$$

The constant Gaussian noise assumption utilized in the proposals can be considered a simplified strategy for computational feasibility.

Remark 2.3. In actual computations, when $\|t_{\bullet g}\|=0$ for some g, this gene will be skipped in the computation. The condition (2.5) is not stringent if the dropout effect is not significant. The normalization $\|t\|=1$ in (2.4) is to ensure the existence of positive rates β and avoid trivial solution as $t,x\to 0$. Another choice $\|x\|=1$ may not guarantee such positive solution. We also note that under such constraint, the obtained gene rescaling parameters β_g and gene-shared latent times t_c are still relative quantities. They are different with the absolute physical kinetic values or physical time up to an unknown global scaling factor, given that only the phase plot of unspliced versus spliced counts are available within scRNA-seq snapshot data. The absolute physical values could be further determined by incorporating experimental measurements such as the metabolic labelling technique [36]. In addition, the key difference between Proposals 2.1 and 2.2 is that they have the following comparative forms:

$$t_{cg}=t_c\beta_g+\beta_g\epsilon_{cg}\ \text{(Proposal 2.1)},\quad t_{cg}=t_c\beta_g+\epsilon_{cg}\ \text{(Proposal 2.2)}.$$

That is why we call Proposal 2.1 the multiplicative noise case, while Proposal 2.2 the additive noise case. As stated in Remark 2.2, it is not clear a priori which choice is more reasonable in practical situations. If the variances of noise are heterogeneous for different genes, the weights should be introduced to take into account the importance of genes in the optimization. However, it is generally infeasible to know this information from the data. In any case, β_g can still be inferred from our proposals to achieve a better consistency of gene-shared latent time t_c.

Both of the above two proposals assume that all of the gene expressions start from a common initial time, which is defined as 0. This assumption may be too strong since the expression for different genes may start from different instants. An extension of this point and general consideration of dropout effect for the gene-shared latent time are studied in [45].

2.3 Numerical validation

To verify the effectiveness of the proposals considered in Section 2.2, we make an illustration with a synthetic example. We simulated 1000 cells with 2000 genes in the on stage by first sampling parameters $(\alpha_g,\beta_g,\gamma_g)$, whose distribution is set to be lognormal(μ,Σ), in which $\mu=[5,0.2,0.05],\Sigma_{11}=\Sigma_{22}=\Sigma_{33}=0.16,\Sigma_{12}=\Sigma_{21}=0.128,\Sigma_{23}=0.032$ (Fig. 2(A)). This results in a typical scale of 100 for the simulated mRNA counts. To avoid the case that the system is almost at steady state and the majority of fluctuations are caused by the observation noise, we sampled the physical real time $t^{(r)}=(t_c^{(r)})_c$ for the cells from a uniform distribution $\mathcal{U}[0,T]$ with T determined as the median of $\tau_g:=2\ln(10)/\beta_g$ for $g=1,\ldots,d$, where the number $2\ln(10)$ in τ_g is chosen such that $u(\tau_g)\approx 0.99\alpha_g/\beta_g$ which is close to the steady state. Then we computed the exact expression number by (2.1), and added a Gaussian noise with mean 0 and standard deviation 30 to form the synthetically measured data (Fig. 2(B)). In the inference stage, we first inferred the parameters by setting the splicing rates $\beta_g=1$, then determined the time-scale parameters by the

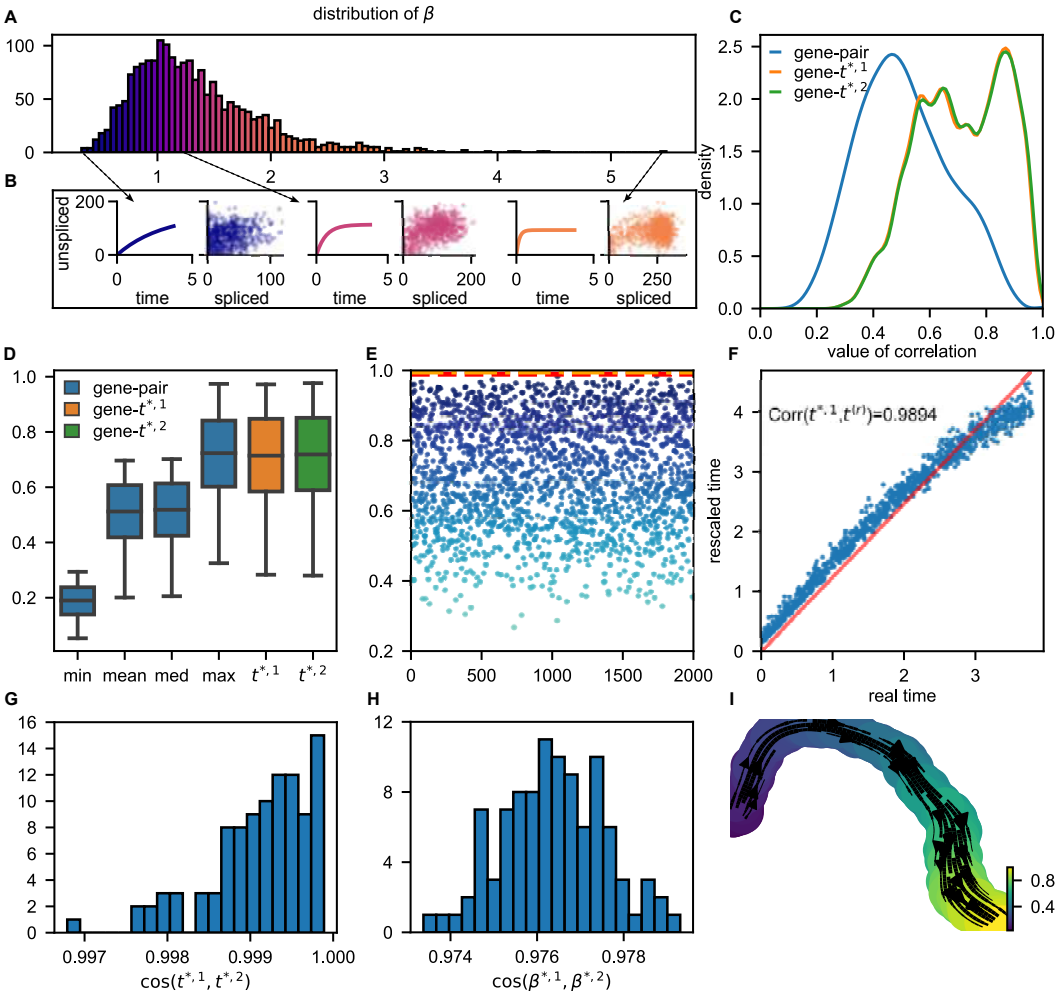

Figure 2: Resolving scaling parameters in simulation data. (A) Empirical distribution of simulated splicing rates β_g, in which the samples are generated from a log-normal distribution. (B) Dynamics of unspliced mRNA and the simulated data with different splicing rates β_g. (C) Comparison between gene-specific time and gene-shared time. The blue curve denotes distribution of the Pearson correlation between each pair of gene-specific time sequences, showing that correlation between direct inferred times is low. The orange and green curves are distribution of correlation between $t^{*,1}$,$t^{*,2}$ and all the gene-specific times, respectively, revealing a more compatible pattern. (D) Further comparison between gene-specific time and gene-shared time. Blue boxes correspond to the distributions of five important statistics of each gene-specific time sequence's correlations. The distribution of correlation between gene-shared time and gene-specific time is drawn as the orange and green boxes. The rescaled time is as consistent as the highest among genes. (E) Correlation of genes-specific time with $t^{(r)}$ compared with correlation between gene-shared time and $t^{(r)}$ (red dashed line for $t^{*,1}$ and orange one for $t^{*,2}$). Rescaled time is the most consistent time sequence to $t^{(r)}$, while some of the gene-specific times highly differ from the ground truth. (F) Scatter plot of $t^{(r)}$ and the rescaled time $t^{*,1}$, the red line denotes the fitted line. The rescaled time shows an obvious linear pattern in regard to the real time. (G) Cosine correlation between $t^{*,1}$ and $t^{*,2}$. The gene-shared times rescaled by two proposals are very close to each other. (H) Cosine correlation between $\beta^{*,1}$ and $\beta^{*,2}$, showing very similar rescaled parameters. (I) Visualization of simulated data and the fitted streamlines based on UMAP [30]. The coloring is based on the rescaled time normalizing to the interval $[0,1]$, which is consistent with the embedded streamlines.

proposed methods in previous subsection. We call the optimized gene-shared common time $t^{*,1}=(t_c^{*,1})_c$ and $t^{*,2}=(t_c^{*,2})_c$ obtained from Proposals 2.1 and 2.2, respectively, and the corresponding optimized splicing rate $\beta^{*,1}$ and $\beta^{*,2}$.

As we cannot recover the physical time $t^{(r)}$of cells due to the scale invariance and an undetermined global timescale, a good rescaling method should improve the linear correlation between the inferred gene-shared time t^* and the gene-specific time $(t_{\bullet g})$. This point is shown in Figs. 2(C) and 2(D), in which we can find that the correlation coefficient distribution and its statistics for the correlation between t^* and $(t_{\bullet g})$ for different g have significant improvements compared with those for the gene pairs (g_1,g_2) (i.e. the correlations $\mathrm{Corr}(t_{\bullet g_1},t_{\bullet g_2})=t_{\bullet g_1}\cdot t_{\bullet g_2}/\|t_{\bullet g_1}\|\|t_{\bullet g_2}\|$).

It is also expected that the inferred gene-shared time t^* has better correlation with the real time $t^{(r)}$ than the gene-pair correlations. This is verified in Fig. 2(E), where we can find that the $(t^{*,1},t^{(r)})$ correlation achieves a high value of 0.9894, which is far bigger than the correlations between gene pairs. Furthermore, the scatter plot of $(t_c^{(r)},t_c^*)$ for different cells in Fig. 2(F) shows an evident linear relation, and this linear dependence is better at early stage of the gene expression, and slightly deteriorates in later stage when the expression reaches steady states. Similar pattern can be also observed in the off stage and we omit it.

To understand the relation between the optimized time obtained from two proposals, we perform 100 times of independent simulations by the same workflow, thus obtain 100 pairs of $(t^{*,1},t^{*,2})$ and $(\beta^{*,1},\beta^{*,2})$. In Figs. 2(G) and 2(H), we present the distribution of correlation coefficients for the optimized gene-shared time t^* and splicing rates β^* from two proposals, respectively. It shows that the two proposals give very close results in terms of the cosine correlations, which are around 0.999 for t^* and 0.976 for β^*. This suggests both options are acceptable choices. In Fig. 2(I), we present the streamline plot of the inferred RNA velocity with the UMAP representation and the smoothed cell coloring according to the gene-shared latent time t^*, which shows nice consistency between the developing flow and time progression of cells.

The issue of being unable to fix the undetermined global time-scale by the proposed approaches is due to the intrinsic drawback of the current experimental techniques. Resolution of this issue depends on further progress of the sequencing technology to extract the temporal information, such as the recent metabolic labeling technique [36].

3 Uncertainty quantification of RNA velocity

In previous section, we proposed a method to unify the time scale between different genes which is critical to the complete parameter inference of RNA velocity models. After parameters are determined, it is also important to evaluate the quality as well as quantify the uncertainty of the inferred parameters and computed RNA velocity. Therefore, we will study the confidence interval construction of RNA velocity models through the Fisher information approach and SEM (supplemented EM) algorithm [32].

3.1 Problem setup

For the observed data $x_{\text{obs}}=(x_{cg})_{cg}=(u_{cg},s_{cg})_{cg}$, we want to maximize the log-likelihood

$$L(\theta|x_{\text{obs}})=\log p(x_{\text{obs}}|\theta)=\log\left[\prod_{cg}\int_{\mathbb{R}}p(x_{cg},t|\theta)\mathrm{d}t\right],$$

where $\theta=(\alpha_g,\beta_g,\gamma_g)_g$, and $p(x,t|\theta)$ is the joint distribution of (x,t) when θ is fixed. The marginal distribution $\int_{\mathbb{R}}p(x,t|\theta)\mathrm{d}t$ is also called the occupancy distribution of cells in [12]. In general $p(x,t|\theta)$ has the form

$$p(x,t|\theta)=p(x|t,\theta)p(t|\theta),$$

where $p(t|\theta)$ is the assumed distribution of the physical time of cells in the considered snapshot data. A working assumption on $p(t|\theta)$ is the natural choice $p(t|\theta)\equiv p(t)=\chi_{[0,T]}(t)/T$, i.e. the uniform distribution on $[0,T]$, which is independent of the parameter θ.

In the inference process, we assume the observation noise is Gaussian with mean 0 and variance σ^2 for all cells and genes. Then, the log-likelihood is

$$L(\theta|x_{\text{obs}})=\log\left[\prod_{cg}\int_0^T\frac{1}{2\pi\sigma^2}\exp\left(-\frac{\|x_{cg}-x_{cg}(t_{cg};\theta_g)\|^2}{2\sigma^2}\right)\cdot\frac{1}{T}\mathrm{d}t_{cg}\right]$$

upon taking the independent-t model discussed in [26]. From the analysis in [26], we know that

$$\begin{aligned}&\log\big(p(x_{cg},t_{cg}|\theta)\big)=-\|x_{cg}-x_{cg}(t_{cg};\theta_g)\|^2+C,\\&p(t_{cg}|x_{cg},\theta)\propto\exp\left(-\frac{\|x_{cg}-x_{cg}(t_{cg};\theta_g)\|^2}{2\sigma^2}\right).\end{aligned}\tag{3.1}$$

Considering t as the latent variable and utilizing the EM algorithm, we have

$$\theta^{(k+1)}=\underset{\theta}{\operatorname{argmin}}\int_0^T\cdots\int_0^T\sum_{cg}\|x_{cg}-x_{cg}(t_{cg};\theta_g)\|^2\exp\left(-\frac{\big\|x_{cg}-x_{cg}\big(t_{cg};\theta_g^{(k)}\big)\big\|^2}{2\sigma^2}\right)\prod_{cg}\mathrm{d}t_{cg}.$$

As $\sigma\to0$, by Laplace asymptotics [2,26], we obtain

$$\text{E-Step:}\quad t_{cg}^{(k)}=\underset{t}{\operatorname{argmin}}\big\|x_{cg}-x_{cg}\big(t;\theta_g^{(k)}\big)\big\|^2,\tag{3.2}$$

$$\text{M-Step:}\quad \theta^{(k+1)}=\underset{\theta}{\operatorname{argmin}}\sum_{cg}\big\|x_{cg}-x_{cg}\big(t_{cg}^{(k)};\theta_g\big)\big\|^2.\tag{3.3}$$

The EM algorithm to infer the parameters is performed by iteratively estimating the rates θ and latent time t through (3.2) and (3.3) until convergence. Our goal in this section is to quantify the uncertainty of the inferred parameters.

3.2 Confidence interval construction through Fisher information

The uncertainty of the maximum likelihood estimator (MLE) can be quantified based on the classical theory of point estimation [24]. For independent and identically distributed data, the MLE $\hat{\theta}_n$ obtained from n samples $\{x_i\}_{i=1:n}$ converges to the true parameter θ^* under suitable regularity conditions on P in the following sense:

$$\sqrt{n}(\hat{\theta}_n-\theta^*) \xrightarrow{d} \mathcal{N}\big(0,I^{-1}(\theta^*)\big) \quad \text{as} \quad n \to \infty, \tag{3.4}$$

where

$$I(\theta)=-\int \nabla_\theta^2 \log p(x|\theta)p(x|\theta)\mathrm{d}x \tag{3.5}$$

is the Fisher information matrix, and the convergence "$\xrightarrow{d}$" holds in the sense of distribution. Hence, for large enough n, the error is approximately normally distributed

$$(\hat{\theta}_n-\theta^*) \overset{d}{\approx} \mathcal{N}\left(0,\frac{I^{-1}(\theta^*)}{n}\right),$$

which means that the $\hat{\theta}_n$ converges to θ^* with the error of magnitude $1/\sqrt{n}$ and a constant characterized by the inverse of the Fisher information matrix at the true value θ^*. In practical computations, the uncertainty of the estimator $\hat{\theta}_n$ can be quantified based on approximating the Fisher information matrix $I(\theta^*)$ by its empirical form

$$\hat{I}\big(\theta^*|x_{\text{obs}}\big) \approx \hat{I}\big(\hat{\theta}_n|x_{\text{obs}}\big) := -\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2 \log p\big(x_i|\hat{\theta}_n\big).$$

For the problem involving latent variables, the calculation of the Fisher information matrix $\hat{I}(\hat{\theta}_n|x_{\text{obs}})$ is not straightforward since the computation of the probability $p(x|\theta)$ involves the integral with respect to the latent time t. Fortunately, this issue has been studied in [32], and the proposed approach can be utilized to approximate $I^{-1}(\theta^*)$ directly.

Following [32], we define the empirical information matrix $\hat{I}_o$ with the observed data x_{obs} as

$$\hat{I}_o(\theta|x_{\text{obs}})=-\frac{1}{n}\nabla_\theta^2 L(\theta|x_{\text{obs}})=-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2 \log p(x_i|\theta),$$

and its inverse at $\theta=\theta^*$ (if the inverse exists)

$$\hat{V}(\theta^*)=\big(\hat{I}_o(\theta^*|x_{\text{obs}})\big)^{-1}.$$

As shown in (3.4), the matrix $\hat{V}^*$ characterizes the uncertainty of the estimated parameter $\hat{\theta}_n$. We can further define the complete-data information matrix with partial observable

$$\hat{I}_{oc}(\theta|x_{\text{obs}},t)=-\frac{1}{n}\nabla_\theta^2 L(\theta\,|\,x_{\text{obs}},t)=-\frac{1}{n}\sum_{i=1}^{n}\nabla_\theta^2 \log p(x_i,t|\theta).$$

It is usually a simple function (e.g. Eq. (3.1)), whose expectation about the conditional distribution $p(t|x_{\text{obs}},\theta)$ evaluated at $\theta=\theta^*$ is

$$\hat{I}_{oc}=\mathbb{E}_{t|x_{\text{obs}},\theta}\big[\hat{I}_{oc}(\theta|x_{\text{obs}},t)\big]\big|_{\theta=\theta^*}=-\frac{1}{n}\sum_{i=1}^{n}\int\nabla_\theta^2\log p\big(x_i,t|\theta^*\big)\,p\big(t|x_i,\theta^*\big)dt.$$

From [32], the EM algorithm (3.2)-(3.3) can be viewed as a mapping $\theta\to M(\theta)$ from the parameter space to itself, which has the form

$$\theta^{(k+1)}=M(\theta^{(k)}),\quad k=0,1,\dots.$$

If $\theta^{(k)}$ converges to θ^* in the parameter space and $M(\theta)$ is continuous, then we have $\theta^*=M(\theta^*)$. By Taylor expansion in the neighborhood of θ^*, we get

$$\theta^{(k+1)}-\theta^*\approx J_M\cdot(\theta^{(k)}-\theta^*),\quad (J_M)_{ij}=\left(\frac{\partial M_i(\theta)}{\partial\theta_j}\right)\Big|_{\theta=\theta^*}.$$

With the formula of total probability

$$p(x_{\text{obs}},t|\theta)=p(x_{\text{obs}}|\theta)p(t|x_{\text{obs}},\theta),$$

we have

$$\log p(x_{\text{obs}}|\theta)=\log p(x_{\text{obs}},t|\theta)-\log p(t|x_{\text{obs}},\theta).\tag{3.6}$$

Taking expectation to both sides of (3.6) with respect to $p(t|x_{\text{obs}},\theta)$, we get

$$\hat{I}_o\big(\theta^*|x_{\text{obs}}\big)=\hat{I}_{oc}-\hat{I}_{om}=\hat{I}_{oc}\big(I-\hat{I}_{oc}^{-1}\hat{I}_{om}\big),$$

where

$$\hat{I}_{om}:=\frac{1}{n}\mathbb{E}_{t|x_{\text{obs}},\theta}\big[-\nabla_\theta^2\log p(t|x_{\text{obs}},\theta)\big]\big|_{\theta=\theta^*}$$

is the missing information. The key observation in [32] is that $J_M=\hat{I}_{oc}^{-1}\hat{I}_{om}$, which is illustrated as below.

Implementation of the EM iterations from $\theta^{(k)}$ to $\theta^{(k+1)}$ is usually performed by taking the maximization of

$$Q(\tilde{\theta}|\theta):=\int L\big(\tilde{\theta}|x_{\text{obs}},t\big)\,p\,(t|x_{\text{obs}},\theta)\,dt$$

with respect to $\tilde{\theta}$, i.e. we have

$$g(\theta^{(k+1)},\theta^{(k)}):=\int\nabla_\theta L\big(\theta^{(k+1)}|x_{\text{obs}},t\big)\,p\big(t|x_{\text{obs}},\theta^{(k)}\big)dt=0.\tag{3.7}$$

Generally denote (3.7) as $g(\tilde{\theta},\theta)=0$ where $\tilde{\theta}=M(\theta)$. We can further take derivative of $g(\tilde{\theta},\theta)$ with respect to θ to obtain

$$\frac{\partial g(\tilde{\theta},\theta)}{\partial\tilde{\theta}}\frac{\partial M}{\partial\theta}+\frac{\partial g(\tilde{\theta},\theta)}{\partial\theta}=0.$$

This leads to

$$\left.\frac{\partial M}{\partial \theta}\right|_{\theta^*} = -\left.\frac{\partial g(\tilde{\theta},\theta)}{\partial \tilde{\theta}}\right|^{-1}_{(\theta^*,\theta^*)} \left.\frac{\partial g(\tilde{\theta},\theta)}{\partial \theta}\right|_{(\theta^*,\theta^*)}. \tag{3.8}$$

Some algebraic manipulations show that

$$\left.\frac{\partial g(\tilde{\theta},\theta)}{\partial \tilde{\theta}}\right|_{(\theta^*,\theta^*)} = -n\hat{I}_{oc}, \quad \left.\frac{\partial g(\tilde{\theta},\theta)}{\partial \theta}\right|_{(\theta^*,\theta^*)} = n\hat{I}_{om}.$$

Substitute these relations into (3.8), we get $J_M = \hat{I}_{oc}^{-1}\hat{I}_{om}$, and the uncertainty covariance matrix

$$\hat{V}(\theta^*) = (I - J_M)^{-1}\hat{I}_{oc}^{-1}.$$

In practical computations, θ^* should be replaced with $\hat{\theta}_n$, i.e. the convergence value of EM iterations. The Jacobian J_M at $\theta = \hat{\theta}_n$ can be approximated by simple difference quotient strategy with a prescribed suitable step size. As shown in [32], we first use a Taylor series expansion to linearize $L(\theta|x_{\text{obs}},t)$ at $t^{(0)}$, then $\hat{I}_{oc}$ is obtained by substituting the conditional expectation of $S(x_{\text{obs}},t)$ found at the last E step of EM for the $S(x_{\text{obs}},t)$ in $\hat{I}_{oc}^{-1}(\theta^*|S(x_{\text{obs}},t))$, where $S(x_{\text{obs}},t)$ is a vector of complete-data sufficient statistics. We present the pseudocode of the overall algorithm in Appendix C.

In most cases, we only care about the diagonal components $\hat{v}_{ii}^*$ of $\hat{V}(\hat{\theta}_n)$ since they are directly related to the variances of the components $\hat{\theta}_{n,i}$ for $i = 1,\ldots,d$. According to (3.4), we have an approximately 95% confidence interval

$$\left(\hat{\theta}_{n,i} - 1.96\sqrt{\frac{\hat{v}_{ii}^*}{n}}, \hat{\theta}_{n,i} + 1.96\sqrt{\frac{\hat{v}_{ii}^*}{n}}\right), \quad i = 1,\ldots,d$$

such that θ_i^* falls in this interval. One important issue is that $\hat{V}^*$ should be positive definite according to its probabilistic meaning. However, it is not guaranteed automatically. Some discussions about this point can be referred to [31, 39].

3.3 Numerical validation

Next, we validate the constructed confidence interval for RNA velocity model using simulation data. We mainly test the accuracy of (α,γ) estimation with different parameter settings under various stages. Steady states of active transcription and inactive silencing can be reached when the induced and repressed transcriptional phases last long enough, respectively. However, these steady states are often difficult to capture, and most processes enter the next life process without reaching the steady state. Here we use models that have not reached the steady state for parameter estimation.

For the on-stage, from [26], we know that the analytical solution of system (2.1) is

$$\begin{aligned} u(t) &= u_0 e^{-\beta t} + \frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right), \\ s(t) &= s_0 e^{-\beta t} + \frac{\alpha}{\beta}\left(1 - e^{-\beta t}\right) - (\alpha - \beta u_0) t e^{-\beta t} \end{aligned} \tag{3.9}$$

for $t < t_s$. In order to verify the sensitivity of different transcription stages to changes in splicing kinetic parameters, we simulated $n = 800$ cells with $d = 20$ genes in the on-stage. For convenience, we chose $\beta = 1$ to avoid scale invariance issue and generate 20 pairs of $(\alpha_g, \gamma_g)_g$ in which $\alpha = 20{:}0.5{:}29.5$ and $\gamma = 1.5{:}0.05{:}2.45$ to test the construction of confidence interval. The physical times of cells were sampled from a uniform distribution $\mathcal{U}[0,T]$ with $T = 2\ln(10)$ to avoid the case that the system reaches and stays at the steady state.

For the off-stage, from [26], we know that the analytical solution of off-stage is

$$\begin{aligned} u(t) &= u_s e^{-\beta(t-t_s)}, \\ s(t) &= s_s e^{-\gamma(t-t_s)} - \frac{\beta u_s}{\gamma-\beta}\left(e^{-\gamma(t-t_s)} - e^{-\beta(t-t_s)}\right) \end{aligned} \tag{3.10}$$

for $t > t_s$. We again simulated 800 cells with 20 genes. The parameters were set to be the same with that in the on stage. The physical times of cells were also sampled from a uniform distribution $\mathcal{U}[0,T]$ with $T = 2\ln(10)$.

The observed data was generated by adding Gaussian noise to the dynamics, that is, $x_{\text{obs}} = x_{\text{true}} + \xi$ with $\xi = \text{normrnd}(\mu,\sigma,2,n)$. In the specific calculation process, u_{obs} and s_{obs} are $n \times d$ matrices, where the observations u^i_{obs} and s^i_{obs} are the elements representing the i-th column in the corresponding matrix, i.e. the unspliced and spliced mRNA for a particular gene, and the u_{obs} and s_{obs} produced have the same variance with $\mu = 0$ and $\sigma = 0.2$. Fig. 3(B) (left panel and middle panel) shows the inference results of parameters α and γ at on-stage, respectively, together with the confidence interval of the parameters. Fig. 3(B) (right panel) shows the inference results of parameter γ. It can be seen from Fig. 3(B) that reasonable inference results can be obtained for Gaussian noise in both on-stage and off-stage.

In order to further validate the constructed confidence interval, we show the distribution of the inference parameters α and γ in Fig. 3(C). To better present the results, we only use 10 genes here whose parameters are $\alpha_g = 21{:}1{:}30, \beta_g = 1$ and $\gamma_g = 1.6{:}0.1{:}2.5$. We added Gaussian noise and performed inference for each pair of parameters for 100 times to show the violin plot. As can be seen from these figures, the estimated parameters closely match the true quantities which shows the reliability of the inference procedure. Also, this result can validate the confidence interval as shown in Fig. 3(B).

In order to show the reliability of the inferred RNA velocity, we compared the norm of the inferred RNA velocity with the norm of the actual velocity, and the cosine value of the angle between the two velocities in Fig. 3(D). We randomly sampled 100 pairs of log-normally distributed parameters (α_g, γ_g), i.e. $\theta = (\alpha,\gamma)$ with $\log(\theta) = N(\mu_1,\Sigma)$, where $\mu_1 = (3, 0.15)$ and $\Sigma = 0.1 I_2$. We assumed that the observation duration $T = 2$ and cell times were randomly sampled from $\mathcal{U}[0,T]$ with noise ξ, i.e. we used 100 genes and 800 cells to infer the RNA velocity. It can be seen from Fig. 3(D) that the ratio of the norms and the cosine value of the velocity angles are distributed around 1, which indicate that our inferred velocity size and direction are reliable.

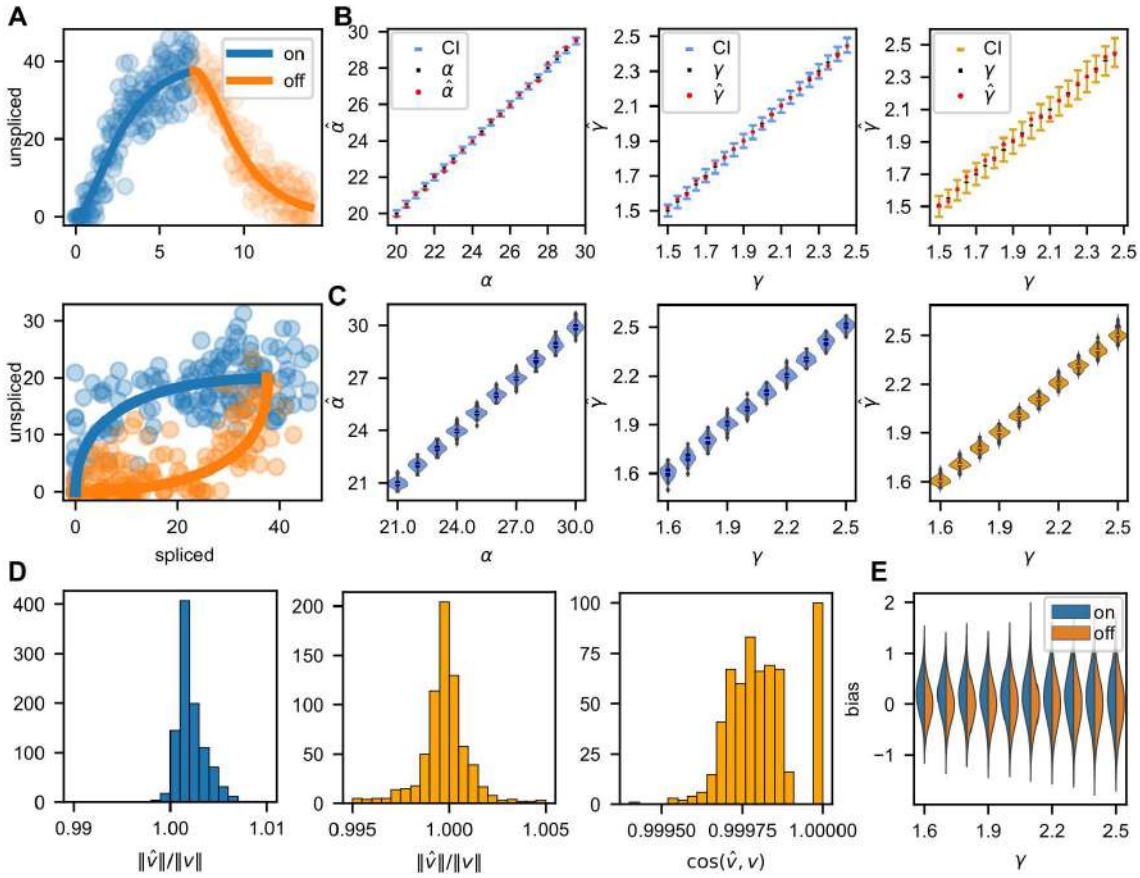

Figure 3: Uncertainty quantification of RNA velocity using simulation data. (A) Simulation of transcriptional process captures transcriptional induction and repression ("on" and "off" stages) of unspliced and spliced mRNA. The top panel shows the abundance of spliced mRNA and the bottom panel shows unspliced and spliced mRNA in phase space. To distinguish the results, we use blue plots for on-stage and orange for off-stage. (B) The 95%-prediction intervals are presented together with the inferred parameters ("on" and "off" stages). It can be found that almost all of the parameters we infer re within the confidence interval. (C) The violin plots depict the distribution of inferred parameters of on-stage (left panel and middle panel) and off-stage (right panel). Fitted parameters mostly lie in a small range. (D) The norm of inferred RNA velocity is compared to the norm of true velocity at on-stage (left panel) and off-stage (middle panel) which show high concentration near 1, also, the cosine of the angle between inferred velocity and true velocity is shown for off-stage (right panel) which also distributed near 1. (E) The bias of inferred velocity under different degradation rates γ at on-stage and off-stage, which is close to normal distribution.

We denote our inferred velocity as $\hat{v}$. Through the definition of RNA velocity $v^* = u - \gamma s$ when assuming $\beta_g = 1$ for all genes, we have

$$\hat{v} \approx u + \xi_1 - \gamma(s + \xi_2) = v^* + (\xi_1 - \gamma\xi_2),$$

which implies that the actual and inferred velocity are not only affected by noise but also related to the selection of γ, thus the ratio of the norm of velocities is affected by γ. In order to demonstrate this statement, we further studied the impact of γ and displayed the

results in Fig. 3(E). Here we simulated 1600 cells with 1000 genes in which half the cells were in the on-stage while others were in the off-stage. As we aim to test the influence of γ, we sampled 100 values of α from a log-normally distribution with $\mu=3$ and $\sigma=0.1$, and constructed 1000 genes with $\gamma=1.6{:}0.1{:}2.5$ paired with sampled α. Then the inference method was applied to the simulated data and we obtained the bias between the true velocity and the inferred velocity. Note that we computed the bias gene-wise here, i.e. for each γ, we tested the bias of inferred velocity under different stages, various cell physical times and values of α. It can be seen from the figure that the error of each selected component is close to normal distribution and the variance increases as γ increases.

4 Optimal choice of kernel bandwidth in random walk construction

Assuming that the RNA velocity obtained accurately describes the actual dynamics locally, a key step in the downstream analysis is the construction of a cellular random walk, i.e. the Markov model on data, which reflects more global and long-time information about cell state-transition dynamics [23, 37, 52], and is also widely used to visualize the streamlines of RNA velocity and the embedding of cells in current practice [1, 3, 21]. In our previous work [26], we derived the continuum limits (i.e. differential equations) of cellular random walk induced by various RNA velocity kernels. Empirically, when dealing with single-cell data of finite or sometimes limited sample size, the hyper-parameters (especially the bandwidth ϵ in Gaussian kernel or number of neighbors k in kNN kernel) in random walk construction have a significant impact on the quantitative behavior of dynamics and downstream analysis. We will study the effects of hyper-parameters and sample size on the random walk convergence rate, elucidating their optimal choice in practice and gaining insights for algorithm implementation. For simplicity, we will only focus on the choice of the optimal kernel bandwidth ϵ for Gaussian-cosine scheme. The other cases can be analyzed similarly.

4.1 Problem setup

Let $(u_i,s_i)\in\mathbb{R}^{2d}$ be the unspliced and spliced gene expression vectors of cell i for $i=1,2,\ldots,n$. To define the probability of transition dynamics between different cells, the randomness introduced by extrinsic or intrinsic noise [8, 18, 51] as well as directed state-transition in relation to RNA velocity needs consideration [3, 21]. The transition between two cells usually involves both drift and diffusion effects. For the diffusion part, the popular Gaussian diffusion kernel has the form

$$d_\epsilon(s_i,s_j)=h\left(\frac{\|s_i-s_j\|^2}{\epsilon}\right),$$

where the function $h(x)$ is usually chosen as a smooth function with exponential decay. For the drift part, we consider the velocity kernel $v(s_i,s_j)=g(\cos\langle\delta_{ij},v_i\rangle)$, where $\delta_{ij}=s_j-s_i$, $v_i=\beta\circ u_i-\gamma\circ s_i$ is the RNA velocity, $\langle\delta_{ij},v_i\rangle$ represents the angle between δ_{ij} and v_i, and $g(\cdot)$ is a bounded, positive, and non-decreasing function. The overall transition kernel is then defined by

$$k_\epsilon(s_i,s_j)=d_\epsilon(s_i,s_j)\cdot v(s_i,s_j).$$

And the transition probability matrix $P_\epsilon=(p_{ij})_{i,j=1:n}$ among cells through the Gaussian-cosine scheme is defined by

$$p_{ij}=\frac{k_\epsilon(s_i,s_j)}{\sum_{j=1}^n k_\epsilon(s_i,s_j)},\quad s_j\sim q(y),$$

where $\sum_{j=1}^n k_\epsilon(s_i,s_j)$ are row normalization factors.

The study of the continuum operator limit of P_ϵ when the number of samples is assumed as infinity has been investigated in [26] by considering the operator $\mathcal{G}_\epsilon$ acting on a smooth function f defined as

$$\mathcal{G}_\epsilon f(x)=\frac{1}{\epsilon^{d/2}}\int_{\mathbb{R}^d}k_\epsilon(x,y)f(y)\mathrm{d}y.$$

From Lemma A.1 in Appendix A (i.e. [26, Lemma 4.1]), the operator $\mathcal{G}_\epsilon$ for Gaussian-cosine scheme has the expansion

$$\begin{aligned}\mathcal{G}_\epsilon f(x)&=\frac{1}{\epsilon^{d/2}}\int k_\epsilon(x,y)f(y)\mathrm{d}y\\&=m_0f(x)+\sqrt{\epsilon}m_1\hat{v}(x)\cdot\nabla f(x)+\mathcal{O}(\epsilon),\end{aligned}\tag{4.1}$$

where m_0,m_1 are constants depending on functions g,h in the diffusion and velocity kernels (see detailed connections in Appendix A), and $\hat{v}(x):=v(x)/\|v(x)\|$ where $v(x)$ is the RNA velocity in the continuum formulation.

Then given the sample probability density $q(\cdot)$, the continuous transition kernel has the form

$$p_\epsilon(x,y)=\frac{k_\epsilon(x,y)q(y)}{d_\epsilon(x)},\quad d_\epsilon(x)=\int k_\epsilon(x,y)q(y)\mathrm{d}y.$$

Define the operator

$$\mathcal{P}_\epsilon f(x)=\int p_\epsilon(x,y)f(y)\mathrm{d}y,$$

and the discrete generator

$$\mathcal{L}_\epsilon=\frac{\mathcal{P}_\epsilon-I}{\sqrt{\epsilon}}.\tag{4.2}$$

From [26, Theorem 4.1], we have the convergence of the generator for the Gaussian-cosine scheme

$$\lim_{\epsilon\to0+}\mathcal{L}_\epsilon f=\mathcal{L}f:=\frac{m_1}{m_0}\hat{v}(x)\cdot\nabla f(x),\quad \hat{v}(x):=\frac{v(x)}{\|v(x)\|}.\tag{4.3}$$

Indeed, we can further identify the higher order expansion of $\mathcal{L}_\epsilon$ as

$$\mathcal{L}_\epsilon f(x) = \mathcal{L} f(x) + \mathcal{O}(\sqrt{\epsilon}), \tag{4.4}$$

since

$$\begin{aligned}\mathcal{P}_\epsilon f(x) &= \frac{\mathcal{G}_\epsilon(fq)(x)}{\mathcal{G}_\epsilon q(x)} = \frac{m_0 f(x)q(x) + \sqrt{\epsilon} m_1 \hat{v}(x)\cdot\nabla(fq)(x) + \mathcal{O}(\epsilon)}{m_0 q(x) + \sqrt{\epsilon} m_1 \hat{v}(x)\cdot\nabla q(x) + \mathcal{O}(\epsilon)} \\ &= f(x) + \sqrt{\epsilon}\mathcal{L} f(x) + \mathcal{O}(\epsilon).\end{aligned}$$

In practical computation, we merely get a limited amount of samples. Beyond investigating the convergence result of the discrete operator to the continuous infinitesimal generator in the limit $\epsilon \to 0$ when the sample size is assumed as infinity, it is also important to understand the optimal kernel bandwidth ϵ when the sample size n is finite. This can be achieved by analyzing the bias and variance tradeoff of discrete models in approximation to their continuum limit.

4.2 Estimation of operator convergence and algorithmic insight

When the sample size n is finite, the discrete generator $\mathcal{L}_{\epsilon,n}$ acting on a smooth function f is defined as

$$\mathcal{L}_{\epsilon,n} f(x) = \frac{1}{\sqrt{\epsilon}}\big(P_{\epsilon,n} f(x) - f(x)\big) = \frac{1}{\sqrt{\epsilon}}\left(\frac{n^{-1}\sum_{j=1}^n k_\epsilon(x,s_j) f(s_j)}{n^{-1}\sum_{j=1}^n k_\epsilon(x,s_j)} - f(x)\right). \tag{4.5}$$

Then, we have the following estimate.

Theorem 4.1 (Finite Sample Approximation of the Operator $\mathcal{L}_\epsilon$)**.** *Let $s_1, s_2, \cdots, s_n$ be n independent and identically distributed samples in $\mathbb{R}^d$ with probability density $q(x)$. Suppose that $f \in C_0^\infty(\mathbb{R}^d)$, which is a smooth function with compact support. Then, we have the error estimate*

$$|\mathcal{L}_{\epsilon,n} f(x) - \mathcal{L}_\epsilon f(x)| = \mathcal{O}\left(\frac{1}{\sqrt{n}\epsilon^{d/4}}\right)$$

in the sense that both the probability

$$p(n,\alpha) := \mathbb{P}\big(|\mathcal{L}_{\epsilon,n} f(x) - \mathcal{L}_\epsilon f(x)| > \alpha\big), \quad 1 - p(n,\alpha)$$

have the $\mathcal{O}(1)$ magnitude in $(0,1)$ only when $\alpha = \mathcal{O}(1/(\sqrt{n}\epsilon^{d/4}))$.

The proof of Theorem 4.1 will be deferred to Appendix B. Based on Theorem 4.1 and Eq. (4.4), we obtain the estimate

$$\begin{aligned}&|\mathcal{L}_{\epsilon,n} f(x) - \mathcal{L} f(x)| \\ &= |\mathcal{L}_{\epsilon,n} f(x) - \mathcal{L}_\epsilon f(x) + \mathcal{L}_\epsilon f(x) - \mathcal{L} f(x)| \\ &\le |\mathcal{L}_\epsilon f(x) - \mathcal{L}_{\epsilon,n} f(x)| + |\mathcal{L}_\epsilon f(x) - \mathcal{L} f(x)| \\ &= \mathcal{O}\left(\frac{1}{\sqrt{n}\epsilon^{d/4}} + \sqrt{\epsilon}\right).\end{aligned} \tag{4.6}$$

Compared with the continuum limit result in [26], the $\mathcal{O}(1/(\sqrt{n}\epsilon^{d/4}))$ term quantifies the influence of finite sample size in cellular random walk.

The above estimation suggests that to achieve the optimal approximation of $\mathcal{L}$ by $\mathcal{L}_{\epsilon,n}$, the best choice of ϵ is

$$\epsilon = \mathcal{O}(n^{-2/(d+2)}), \tag{4.7}$$

when the sample size n is fixed. This is obtained when the two error terms in (4.6), the variance and bias, are balanced. In this case, the optimal error of the operator approximation is $\mathcal{O}(n^{-1/(d+2)})$.

The result also shows the curse of dimensionality when using the velocity-induced cellular random walk to approximate the dynamics of continuous ODEs. As the dimensions of the input data d increases, the overall error $\mathcal{O}(n^{-1/(d+2)})$ deteriorates, which suggests insufficient model accuracy. This analysis leads to several possible interpretations for the downstream RNA velocity analysis: The first insight is that to achieve a more stable inference on the cellular development path, we should better use a relatively low dimensional space for genes instead of a very high dimensional space. The second possible interpretation is that although the operator approximation is bad in a very high dimensional space, the overall direction of developments, such as the streamline visualization of the cells along the main backbone, can still be estimated in a relatively accurate manner. The biological pathways and highly correlated gene functional modules underlying real scRNA-seq data [7] could also further reduce the effective dimension of the data manifold, and lead to larger convergence rate empirically than theoretical results [40]. Indeed, the following remark indicates that our analysis in this section can be extended to the manifold case, which is also believed to be relevant for scRNA-seq data points.

Remark 4.1. We can follow the analysis approach in [22] to build the Theorem 4.1 on manifolds. Assume that M is a smooth manifold of dimension m embedded in $\mathbb{R}^d, m<d$, and μ is a probability measure on M, which has a density with respect to the Riemannian measure dx on M (i.e. $d\mu(x)=q(x)dx$). The function $q(x)$ characterizes the density of available sample points. Let $\{e_1,\dots,e_m\}$ be an orthonormal basis of the tangent space $T_xM\subset\mathbb{R}^d$ at x. For any $y\in M$ in the neighborhood of x, we map y to $u\in T_xM$ through an orthogonal projection, and denote $u=(u_1,\dots,u_m)$ the local coordinates of y in terms of the basis $\{e_1,\dots,e_m\}$. Let $(s_1,\dots,s_m)$ be the normal coordinates of y.

We consider the integral (4.1) on manifold M with density function $q(x)$

$$\mathcal{G}_\epsilon f(x)=\frac{1}{\epsilon^{m/2}}\int_M k_\epsilon(x,y)q(y)f(y)\mathrm{d}y.$$

Without loss of generality, we assume $x=0$. Similar to the analysis in Euclidean space, due to the exponential decay of function h, we can restrict the integration to a Euclidean ball of radius $C\sqrt{\varepsilon}$ on M, i.e.

$$\int_M h\left(\frac{\|y\|^2}{\varepsilon}\right)g(\cos\langle y,v\rangle)f(y)q(y)\mathrm{d}y\simeq\int_{\|y\|<C\sqrt{\varepsilon}}h\left(\frac{\|y\|^2}{\varepsilon}\right)g(\cos\langle y,v\rangle)f(y)q(y)\mathrm{d}y.$$

Generally, it is not convenient to calculate the integral in normal coordinates, and we convert it to the tangent space T_xM. By the Taylor expansion of fq around $x=0$ on M, the variable transformation from s to u and its expansion around $u=0$, we have

$$\varepsilon^{\frac{m}{2}}\mathcal{G}_\varepsilon f(0)=\int_{\|u\|<C\sqrt{\varepsilon}}h\left(\frac{\|u\|^2}{\varepsilon}\right)g(\cos\langle u,v\rangle)\left(f(0)q(0)+\sum_{i=1}^m u_i\frac{\partial(fq)}{\partial s_i}(0)\right)\times\left(1+2\sum_{i=1}^m a_i^2u_i^2\right)du+\mathcal{O}(\varepsilon),$$

where a_i is the curvature of the coordinate geodesic along e_i at $x=0$. Then, the analysis and derivations will be conducted in $\mathbb{R}^m$, and we can obtain similar conclusion to those in $\mathbb{R}^d$ above.

4.3 Numerical validation

In this subsection, we present a toy example to show the convergence rate of $\mathcal{L}_{\epsilon,n}f$. We took $d=3$ and chose a linear function $f_1(x)=x_1+x_2+x_3$ and a nonlinear function $f_2(x)=x_3^2$ to perform the numerical simulations. We first generated $n=2000$ samples $(u^{(k)},s^{(k)})_{k=1:n}=(u(t_k),s(t_k))_{k=1:n}$ according to the RNA velocity dynamics

$$\frac{du_g}{dt}=\alpha_g-\beta_g u_g,\quad \frac{ds_g}{dt}=\beta_g u_g(t)-\gamma_g s_g(t),\quad (u_g,s_g)|_{t=0}=(0,0),\quad g=1,\dots,d$$

by choosing $t_k\sim\mathcal{U}[0,T]$ for $k=1,\dots,n$ where $T=2\ln 10$. Here we chose the parameters $\alpha=(20,20.5,21)^\top$, $\beta=(1,1,1)^\top$, and $\gamma=(1.5,1.55,1.6)^\top$, where each component corresponds to the index $g=1,2,3$, respectively. We then generated $n=2000$ samples $\{x_k\}$ with velocity $\{v_k\}$ by setting

$$x_k=s^{(k)}+\epsilon_k,\quad \epsilon_k\sim N(0,0.5I_3),$$

and $v_k=\beta\circ u^{(k)}-\gamma\circ x_k$ for $k=1,\dots,n$. In the downstream analysis, we chose $g(x)=\exp(x)$, $h(x)=\exp(-x)$ and defined the root-mean-squared error as

$$\text{error}=\left[\frac{1}{n}\sum_{k=1}^n\left(\mathcal{L}_{\epsilon,n}f(x_k)-\mathcal{L}f(x_k)\right)^2\right]^{\frac{1}{2}}$$

by averaging over $n=2000$ samples. In the simulation, we chose $\epsilon=0.002:0.002:0.082$ and the results are shown in Fig. 4.

According to the estimate (4.6), we know that when $\epsilon\lesssim\mathcal{O}(n^{-2/5})$, the "variance" term is dominant with the order $1/(\sqrt{n}\epsilon^{3/4})$. Thus, as ϵ increases, the $\ln(\text{error})$ versus $\ln(\epsilon)$ plot should present a linear relation with slope $-3/4$ theoretically. This is verified in Figs. 4(A) and 4(B), in which the linear fitting gives the slope -0.74 for the linear case shown in the left panel and -0.76 for the nonlinear case shown in the right panel. When $\epsilon\gtrsim\mathcal{O}(n^{-2/5})$, the bias term is dominant and the error curve demonstrates a turn-over at $\ln(\epsilon)\sim\ln(n^{-2/5})\approx-3.04$ theoretically, which is close to the computed minimum point at $\ln(\epsilon)\approx-3.45$ in linear case and $\ln(\epsilon)\approx-3.20$ in nonlinear case.

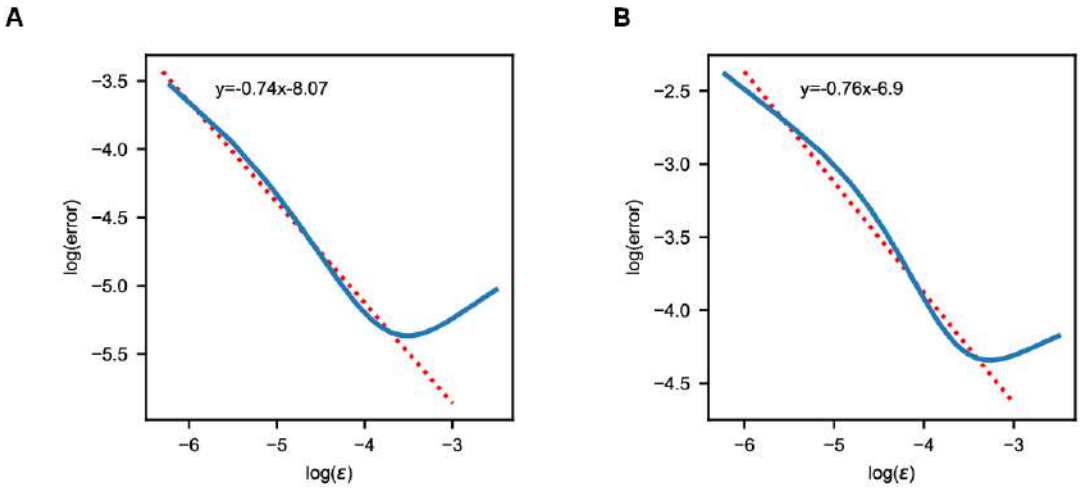

Figure 4: Effect of kernel bandwidth ϵ on operator approximation. The logarithmic plots of the operator approximation error for the linear case $f_1(x)=x_1+x_2+x_3$ (left panel) and nonlinear case $f_2(x)=x_3^2$ (right panel). The errors of both cases are averaged over $n=2000$ samples. The minimal errors are obtained when $\ln(\epsilon)\approx -3.45$ (left panel) and $\ln(\epsilon)\approx -3.20$ (right panel). When $\epsilon \lesssim n^{-2/5}$, the dominant term of the error should be $\mathcal{O}(1/(\sqrt{n}\epsilon^{3/4}))$. This gives the slope $-3/4$ in theory, which is close to the estimated value -0.74 (left panel) or -0.76 (right panel) by linear regression.

5 Transition time estimation among cell states

After constructing the Markov chain among individual cells induced by RNA velocity, the downstream analysis could be performed to reflect the long-term and global (i.e. among multiple cell states) dynamics of cell-state transitions [23, 26, 36]. In our previous work [26, 52], we proposed the approach based on transition path theory to infer coarse-grained cell lineage and quantify corresponding likelihoods from the cellular random walk. Another practical task is to quantify the duration of the transition from one cell to another along the transition paths, i.e. defining the pseudo-temporal distance [43] between cells induced by RNA velocity, which could be realized by the first hitting time analysis described below.

5.1 Problem setup

Suppose that the RNA velocity induces the cellular random walk with transition probability matrix $P=(p_{ij})$ which was defined in Section 4.1. Our goal is to define a pseudo-temporal distance T_i^A from cell i to the cell set A which reflects the state-transition time based on the Markov chain model. When the cell set $A=\{j\}$, T_i^A gives the pseudo evolution time from cell i to cell j.

For Markov chain $\{X_n, n\geq 0\}$, the first hitting time of a set A is defined as

$$\tau^A=\inf\{n\geq 0: X_n\in A\},$$

where A is a subset of the state space. The mean first hitting time for the process to reach A starting from i is given by

$$k_i^A = \mathbb{E}_i(\tau^A) = \sum_{n<\infty} n\mathbb{P}_i(\tau^A = n) + \infty\mathbb{P}_i(\tau^A = \infty),$$

where $\mathbb{E}_i$ and $\mathbb{P}_i$ denotes the expectation and probability conditioned on $X_0 = i$, respectively. The quantity k_i^A serves as the rational proposal for pseudo-temporal distance T_i^A, and we will demonstrate the equations to calculate it below.

We will consider two types of hitting times to describe the evolution time from one cell to another cell set A. Firstly, we show that it is straightforward to use the Eq. (5.1) below to compute the mean first hitting time when there is no bifurcation. Secondly, we use a simplified model to demonstrate the limitation of this approach when there is the "bottleneck" state and cell-state differentiation in dataset. We then show how to get a biologically meaningful time by utilizing the taboo set concept.

5.2 Transition time estimation through first hitting time analysis

The computation of k_i^A is based on the following lemma (see, e.g. [33, Theorem 1.3.5]).

Lemma 5.1. *The vector of mean first hitting times $k^A = (k_i^A)_i$ is the minimal non-negative solution to the system of linear equations*

$$\begin{cases} k_i^A = 0, & i \in A, \\ k_i^A = 1 + \sum_{j \notin A} p_{ij} k_j^A, & i \notin A. \end{cases} \tag{5.1}$$

To solve (5.1), we use the following iterations:

$$K_n^A = \mathbf{1} + QK_{n-1}^A, \quad K_0^A = \mathbf{1}, \tag{5.2}$$

where K_n^A is the n-th iteration of k^A, and $Q = (q_{ij}) := (p_{ij})_{i,j \in A^c}$, i.e. the matrix formed by removing the row and column elements corresponding to $i \in A$ from the transition probability matrix P. Next we show that the iteration (5.2) is a contraction mapping, i.e. the spectral radius $\rho(Q) < 1$.

Theorem 5.1. *Assume the velocity-induced cellular random walk with transition probability matrix P is irreducible, then the iteration* (5.2) *is a contraction mapping, i.e. $\rho(Q) < 1$.*

Proof. Without loss of generality, assume

$$P = \begin{bmatrix} Q & R_1 \\ R_2 & B \end{bmatrix},$$

where Q and B describe the transition probabilities among the states in A^c and A, respectively. R_1 and R_2 describe the transition probabilities from the states in A^c to A and A to A^c, respectively, which should not be zero since P is irreducible.

We will first consider the case that Q is irreducible. In this case, if all of the row sums of the matrix Q are strictly less than 1, then the conclusion holds simply by Gershgorin circle theorem [19]. Otherwise, we have $\rho(Q)\leq 1$ and there is at least one row of Q such that its sum is strictly less than 1. Below we show that the assumption $\rho(Q)=1$ will lead to contradiction.

Utilizing the Perron-Frobenius theorem, we have the Perron vector x with positive components such that

$$Qx=\rho(Q)x=x.$$

Suppose $x_l=\max\{x_k\}_{k\in A^c}$ and define $y=x/x_l$. We have $Qy=y$ and

$$\sum_k q_{lk}y_k=y_l=1.$$

Since $y_k\leq 1$ and $\sum_k q_{lk}\leq 1$, the above identity requires that $y_k=1$ for $q_{lk}>0$, i.e. the neighborhood states of l. We can apply similar arguments to these states k, which eventually lead to $y\equiv 1$. While this contracts with the condition that at least one row rum of Q is strictly less than 1. So, we have $\rho(Q)<1$ when Q is irreducible.

In general cases, we can decompose A^c into several irreducible components and transient states. For each irreducible component, it has the transition probability sub-matrix with similar structure as the above case. Thus, the spectral radius is strictly less than 1. For the transient states, the transition probability sub-matrix has the property that all row sums are less than 1, thus the spectral radius is also strictly less than 1. Overall, we have $\rho(Q)<1$ in the general cases. □

Remark 5.1. The result $\rho(Q)<1$ ensures that $N=(I-Q)^{-1}$ is well-defined, which is called the fundamental matrix in the literature [20]. Here we give a self-contained linear algebra proof instead of probabilistic arguments. The above theorem also tells us that $k^A=N\cdot\mathbf{1}$ although it is not a feasible approach to compute k^A due to the ill-conditioning of $I-Q$.

The above approach is only useful when there are no bifurcations, i.e. no differentiations in cell development. This can be illustrated by a simplified model shown in Fig. 5(A), in which we model the stem cell as state S, the developmental bottleneck as state B, and two differentiated states as C and D, denoting different fates of cell differentiation. The state S can only transit to B, while B is able to transit to C and D, or back to S. We assume the transitions have some preferred directionality, i.e. it is nearly impossible to transit back along the directed developmental pathway. The above assumption amounts to set that p_{SB},p_{BC},p_{BD} are $\mathcal{O}(1)$ and p_{BS},p_{CB},p_{DB} are $\mathcal{O}(\epsilon)$. Heuristically, we can set the transition probability matrix P as

$$P=\begin{pmatrix}0 & 1 & 0 & 0\\ \epsilon & 0 & p & q\\ 0 & \epsilon & 1-\epsilon & 0\\ 0 & \epsilon & 0 & 1-\epsilon\end{pmatrix},$$

in which p and q are probabilities of $\mathcal{O}(1)$ and $p+q=1-\epsilon$.

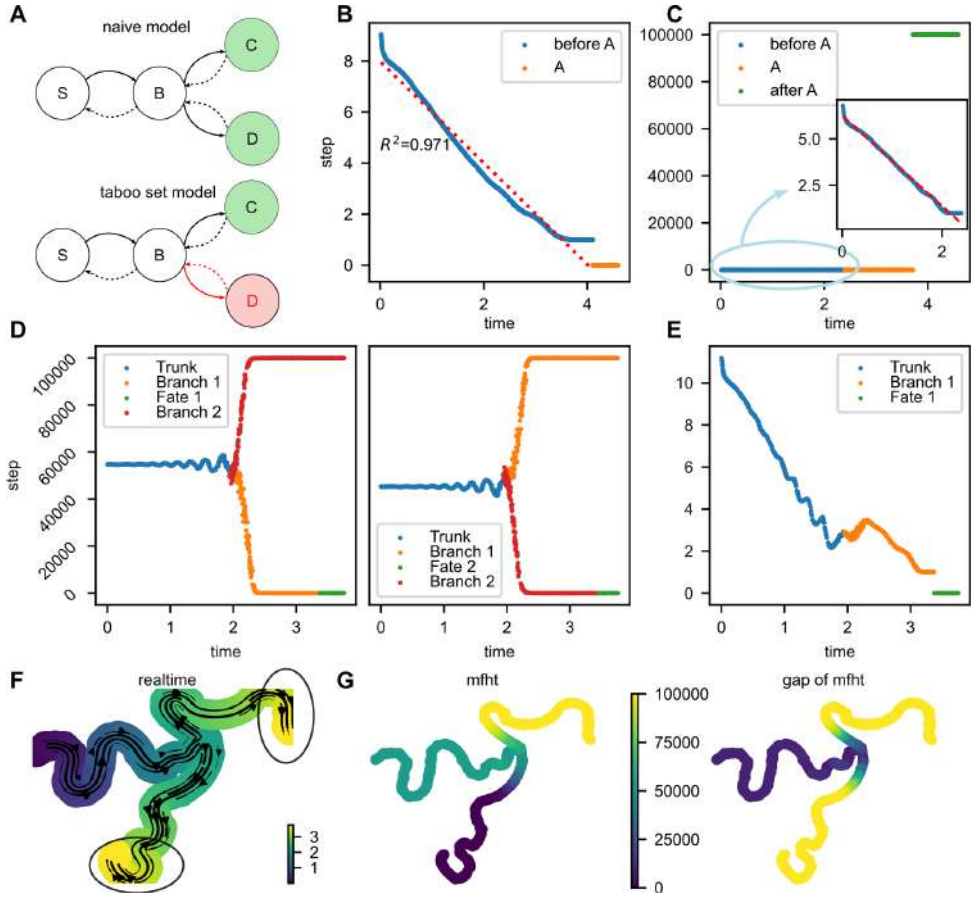

Figure 5: Estimating the pseudo-temporal distance via the first hitting times. (A) Schematics of the naive and taboo set models. (B) The physical time of the synthetic model compared with the mean first hitting time to the target set in the end. There is a linear pattern for the computed mean first hitting time and the red line is obtained by linear regression. (C) The physical time compared with the mean first hitting time to a target set in the middle. For cells before this target set A, there is still a linear pattern which is shown in the inset, in which the red dashed line is obtained by linear regression. (D) In the bifurcation case, the physical time compared with the mean first hitting time to one branch computed by iterative method. Cells in the other branch and before bifurcation have very large mean first hitting times. (E) By setting the other branch as the taboo set, the mean first hitting time shows a nearly linear pattern. (F) Streamline embedded UMAP plot of the synthetic bifurcation data. The two expected termination sets are circled. (G) UMAP of the synthetic data. The left panel is colored according to the mean first hitting time to the lower branch termination cells, and the right panel is colored according to the absolute value of the difference between two computed mean first hitting times to two circled branches in Fig. 5(F).

In this setup, the mean first hitting time of each state to the target set C can be obtained according to (5.1) as

$$\begin{cases} k_S^C = 1 + k_B^C, \\ k_B^C = 1 + q k_D^C + \epsilon k_S^C, \\ k_D^C = 1 + \epsilon k_B^C + (1-\epsilon) k_D^C, \end{cases} \tag{5.3}$$

from which we get

$$k_S^C = 1+\frac{q+\epsilon+\epsilon^2}{\epsilon p}, \quad k_B^C = \frac{q+\epsilon+\epsilon^2}{\epsilon p}, \quad k_D^C = \frac{1+\epsilon^2}{\epsilon p}.$$

We have $k_S^C, k_B^C, k_D^C \sim \mathcal{O}(\epsilon^{-1})$, i.e. we need a prominent long transition time to reach C from any other state, including S and B, which looks counter-intuitive. The reason is that although the states S, B are in the upstream of C in the cell development, they have $\mathcal{O}(1)$ probability to reach D in an $\mathcal{O}(1)$ timescale, while it is very difficult to transit back from D to B once D is reached. This effect finally makes the overall transition time from any other states to C are extremely large. However, this does not reflect the biological intuition that the transition time from S and B to a specific differentiated state C or D is in $\mathcal{O}(1)$ timescale.

Biologically, we are mainly interested in the transition time that the stem/bottleneck cell differentiates to a specific cell type instead of letting it transit to another differentiated state and then get back. In this case, the other differentiated cell states beyond the interested states should be neglected and form a forbidden set. We call this set a taboo set, and compute the mean first hitting time of state S or B to C conditional on not reaching the taboo set $H=\{D\}$. This is illustrated in the taboo set model in Fig. 5(A). Denote the first hitting time with taboo set H by

$${}_H\tau^A = \inf\{n \geq 0 : X_n \in A,\ X_m \notin H \text{ for } m \leq n\},$$

and the mean first hitting time by ${}_Hk^A$,

$${}_Hk_i^A = \sum_{n<\infty} n\mathbb{P}_i({}_H\tau^A = n) + \infty\mathbb{P}_i({}_H\tau^A = \infty).$$

Then ${}_Hk^A$ satisfies

$${}_Hk_i^A = 1+\sum_{j\notin A\cup H} p_{ij}\,{}_Hk_j^A, \quad i\notin A\cup H. \tag{5.4}$$

So we obtain

$$\begin{cases} {}_Hk_S^C = 1+{}_Hk_B^C, \\ {}_Hk_B^C = 1+\epsilon\,{}_Hk_S^C, \end{cases}$$

from which we get

$${}_Hk_S^C = \frac{2}{1-\epsilon} \approx 2, \quad {}_Hk_B^C = \frac{1+\epsilon}{1-\epsilon} \approx 1.$$

This result reflects the intuition that the transition time from S to C and from B to C are about 2 and 1, respectively, by simply counting the transition steps in the Markov chain. By setting the taboo set, the behavior of the tabooed process is similar to the case when there is no bifurcation, and the taboo set model acts as a pruning strategy.

5.3 Numerical validation

To verify the applicability of our proposal on the evolution time estimation, for the non-bifurcation situation, we simulated 1000 cells with 2000 genes at on-stage to generate the synthetic data. The splicing rate β_g was fixed to 1 to avoid considering the scale invariance issue, and the transcription rates α_g and degradation rates γ_g were sampled from a log-normal distribution with mean $\mu=[5,0.05]$ and covariance matrix $\Sigma=0.16I_2$. The physical time of cells were sampled from $\mathcal{U}[0,T]$ with $T=2\ln(10)$. After inferring the parameters, a Gaussian-cosine kernel was constructed. We chose the set A as the top 100 cells having the largest sampled real time, and compute the mean first hitting time from any cell to this set by solving (5.1). From Fig. 5(B) we can find that the computed mean first hitting time matches well with the real time of cells upon ignoring a scaling constant and the R-squared of linear regression is 0.968, which validates our proposal in the considered simple synthetic example.

To compute the mean first hitting time of non-terminal states, we applied the iteration (5.2), from which we can tell the relevant order of cells. We chose the cells with the rank order $500\sim800$ as set A, by which the other cells were separated as two groups with weak links. By iterating for a sufficiently long time, we can tell from the result the relevant order of cells to set A. Shown in Fig. 5(C) is the iterated mean first hitting time after 1×10^5 iterations, as cells after set A are nearly impossible to transit back to cells prior to set A, the mean first hitting time of cells posterior to A is very large and is the scale of iteration time. Mean first hitting times of cells prior to set A still show a linear decay pattern. However, for the cells that are close to A, the transition times show fluctuations and an increasing trend since these cells have small probability to perform transitions to the cells posterior to A due to the weak link between them. This phenomenon also partially suggests the adoption of the taboo set model in this simple case.

To test the taboo set model, we first applied the iterative method directly to a synthetic bifurcation data. Here to produce bifurcation, we sampled parameters $(\alpha_g,\beta_g,\gamma_g)$, whose distribution was set to be lognormal(μ,Σ), in which $\mu=[5,0.2,0.05]$, $\Sigma_{11}=\Sigma_{22}=\Sigma_{33}=0.16$, $\Sigma_{12}=\Sigma_{21}=0.128$, and $\Sigma_{23}=0.032$ which is the same as the setup in Section 2. Then we used the true parameters in the computation of the mean first hitting time. The physical time of cells were sampled from $\mathcal{U}[0,T]$ with T determined as the median of $\tau_g:=2\ln(10)/\beta_g$ for $g=1,\ldots,d$. The bifurcation was produced by adding the switch of gene expression from the on-stage to off-stage. For the first branch, we assigned 70% of the genes to switch to off stage at $2\ln(10)/\beta_g$ and for the second branch, the rest 30% genes are assigned. From Fig. 5(D) we can find that when setting the target set as the termination part of one branch, the mean first hitting time from another branch rapidly grows to a huge number, and the cells before the bifurcation point also have long mean first hitting times, which is similar to our analysis of the simplified naive model in previous subsection. This result implicitly indicates that we can use the naive model to detect where the bifurcation happens. As shown in Fig. 5(E), by setting the second branch to be the taboo set, the mean first hitting time shows a nearly linear pattern both before and after the bifurcation

point, showing the taboo set model can successfully give an estimation of transition time without considering the other bifurcation branches.

The results above show that we can use the mean first hitting time as an estimation of the physical time of cells. Such an approach for pseudo-time can be extended to real-world data in which the ground truth (i.e. physical time, bifurcation lineage, and especially taboo set) is not available. Here we will present a brief idea of the implementation in scRNA-seq data and leave further analysis and software development for future work [45].

From low-dimensional visualization and velocity streamline embedding, we can identify the main fates (or branches) of differentiation as shown in Fig. 5(F) using existing lineage inference methods. To identify the taboo set according to the computational results, we first apply the iterative method to calculate the mean first hitting time to the expected fates on different branches and then take a postprocessing step. As shown in Fig. 5(G), the left panel is colored according to the mean first hitting time to the lower branch termination cells, and the right panel is colored according to the absolute value of the difference between two computed mean first hitting times to two circled branches in Fig. 5(F), which is nearly the iteration number on the two branches while much smaller for cells on the main trunk. Thus, the iterated mean first hitting time could distinguish cells at the branches or at the main trunk. If the ground truth is unknown, cells with large mean first hitting times as well as a significant gap between mean first hitting time to different fates (see the upper branch in Fig. 5(G)) could provide a reasonable candidate of the taboo set. This strategy works for the current synthetic example, however, its application to more practical examples needs more testing and deserves further study in the future.

6 Discussion and conclusion

The RNA velocity analysis has provided useful tools to predict future cell states within snapshot scRNA-seq data by modeling mRNA expression and splicing processes. Despite the established workflow and existing theoretical studies, several issues involved in parameter inference and downstream dynamical analysis of the current RNA velocity model remain elucidated, especially regarding the rationale and robust algorithm design and implementation. In this paper, we proposed several strategies to address these challenges through mathematical or statistical models as well as numerical analysis.

To unify the timescale of RNA velocity dynamics across different genes, we formulated the optimization framework based on either additive or multiplicative noise assumption to optimally determine the gene-specific rescaling parameters and proposed the numerical scheme to efficiently calculate the gene-shared latent time. Beyond the current independent gene assumption, it is possible to extend the current framework to RNA velocity models incorporating gene regulation or interactions [4, 25, 46, 49].

To estimate the uncertainty of the inferred parameters and corresponding RNA velocity, we performed the confidence interval analysis of RNA velocity utilizing Fisher infor-

mation approach, and SEM method [32] was applied to obtain the asymptotic covariance of kinetic parameters in the dynamical RNA velocity model where latent cell time was involved in EM inference. In addition to the deterministic ODE model, our uncertainty quantification analysis and confidence interval construction approach could also be applied to stochastic RNA velocity model based on chemical reactions [13, 14, 26, 46].

To determine the optimal hyper-parameters in the velocity-induced random walk, we analyzed its convergence rate toward continuous ODE dynamics in the operator sense, and assessed the dependence of convergence rate on sample size n and data kernel bandwidth ϵ. The results suggest that choosing kernel bandwidth ϵ around the scale of $\mathcal{O}(n^{-2/(d+2)})$ provides the best operator approximation. It also indicates that as the dimensionality d of the system increases, the accuracy of approximation would be impaired. Consequently, feature selection or dimensionality reduction could improve the cellular random walk construction. In parallel to the random walk approach, another strategy for downstream RNA velocity analysis is to fit the continuous velocity field [36] using vector-valued kernel methods [28] as proposed in Dynamo [36] or neural-ODE methods [5, 6, 27], where similar convergence analysis could also be informative for the algorithm implementation.

To perform lineage inference and assign pseudo-time that is consistent with RNA velocity dynamics, we proposed to use the mean first hitting time of the velocity-induced random walk. The hitting time has been proposed for single-cell lineage analysis by defining lazy-teleporting random walk on cellular similarity graph [41]. In the scVelo package, a velocity pseudotime is defined based on the diffusion-like distance [16, 48] through the eigendecomposition of the weighted cellular velocity graph. In bifurcation systems with strong directionality induced by RNA velocity, our analysis and numerical examples suggest the introduction of taboo set for mean first hitting time analysis. Theoretically, the hitting time has close relation with the commute distance for undirected graph [44] $\mathcal{C}(x_i,x_j)=(T_i^j+T_j^i)^{1/2}$ where T_i^j denotes the mean first hitting time from x_i to x_j. It will be insightful to study the limit of mean first hitting time for the velocity-induced random walk when the sample tends to infinity, as studied for random geometric graph case [44] and unweighted directed graph case [17].

Overall, the numerical analysis and statistical models presented in the current work could serve as a mathematical step towards more robust and effective algorithmic implementation of the RNA velocity model computation and analysis.

Appendix A. Proof of some lemmas

Lemma A.1 (Expansion of the Un-Normalized Kernel k_ϵ)**.** *The operator $\mathcal{G}_\epsilon$ for Gaussian-cosine scheme has the expansion*

$$\begin{aligned}\mathcal{G}_\epsilon f(x) &= \frac{1}{\epsilon^{d/2}}\int k_\epsilon(x,y)f(y)\mathrm{d}y \\ &= m_0 f(x)+\sqrt{\epsilon}m_1\mathcal{A}f(x)+\mathcal{O}(\epsilon),\end{aligned}$$

where

$$\begin{aligned} m_0 &:= \mathcal{G}_\epsilon 1 = \frac{1}{\epsilon^{d/2}} \int k_\epsilon(x,y)\mathrm{d}y \\ &= C_d \int_0^\infty r^{d-1} h(r^2)\mathrm{d}r \int_{-\pi}^{\pi} |\sin\theta|^{d-2} g(\cos\theta)\mathrm{d}\theta, \\ m_1 &:= C_d \int_0^\infty r^{d} h(r^2)\mathrm{d}r \int_{-\pi}^{\pi} \cos\theta|\sin\theta|^{d-2} g(\cos\theta)\mathrm{d}\theta, \end{aligned}$$

and

$$\mathcal{A}f(x) = \|\nabla f(x)\| \cos\langle v(x), \nabla f(x)\rangle = \hat{v}(x)\cdot\nabla f(x), \quad \hat{v}(x) := \frac{v}{\|v\|}.$$

Here, $d>1$,

$$C_d = S_d \left(\int_{-\pi}^{\pi} |\sin\theta|^{d-2}\, \mathrm{d}\theta \right)^{-1},$$

and S_d is the surface area of the d-dimensional unit sphere.

The above lemma is in fact the [26, Lemma 4.1] except that the remainder term is explicitly characterized as $\mathcal{O}(\epsilon)$ instead of $o(\sqrt{\epsilon})$. The proof is by straightforward derivations, which is also shown in Lemma A.2 below.

To get the variance error, we first study the operator $\tilde{\mathcal{G}}_\epsilon$ defined by

$$\tilde{\mathcal{G}}_\epsilon f^2(x) = \frac{1}{\epsilon^{d/2}} \int_{\mathbb{R}^d} \big(k_\epsilon(x,y) f(y)\big)^2 q(y)\mathrm{d}y.$$

The following lemma can be obtained.

Lemma A.2 (Expansion of the Kernel k_ϵ^2)**.** *The operator $\tilde{\mathcal{G}}_\epsilon$ for Gaussian-cosine scheme has the expansion*

$$\begin{aligned} \tilde{\mathcal{G}}_\epsilon f^2(x) &= \frac{1}{\epsilon^{d/2}} \int k_\epsilon^2(x,y) f^2(y) q(y)\mathrm{d}y \\ &= \tilde{m}_0 f^2(x) q(x) + \sqrt{\epsilon}\tilde{m}_1 \mathcal{A}\big(f^2(x)q(x)\big) + \mathcal{O}(\epsilon), \end{aligned}$$

where

$$\begin{aligned} \tilde{m}_0 &= C_d \int_0^\infty r^{d-1} h^2(r^2)\mathrm{d}r \int_{-\pi}^{\pi} |\sin\theta|^{d-2} g^2(\cos\theta)\mathrm{d}\theta, \\ \tilde{m}_1 &= C_d \int_0^\infty r^{d} h^2(r^2)\mathrm{d}r \int_{-\pi}^{\pi} \cos\theta|\sin\theta|^{d-2} g^2(\cos\theta)\mathrm{d}\theta. \end{aligned}$$

Here, $d>1$,

$$C_d = S_d \left(\int_{-\pi}^{\pi} |\sin\theta|^{d-2}\, \mathrm{d}\theta \right)^{-1},$$

and S_d is the surface area of the d-dimensional unit sphere.

Proof. For convenience, let $v(x)=\|v\|(1,0,\cdots,0)^\top$ without loss of generality. Let us first consider the case of $d=2$. Consider 2-dimensional polar coordinates transformation

$$\begin{cases} y_1=x_1+r\cos\theta, \\ y_2=x_2+r\sin\theta, \end{cases}$$

where θ is the angle between $y-x$ and $v(x)$. Then we have

$$\begin{aligned} &\frac{1}{\epsilon}\int \big(k_\epsilon(x,y)f(y)\big)^2 q(y)\mathrm{d}y \\ =&\frac{1}{\epsilon}\int_0^\infty\int_{-\pi}^{\pi} rh^2\left(\frac{r^2}{\epsilon}\right)g^2(\cos\theta)f^2(r,\theta)q(r,\theta)\mathrm{d}\theta\mathrm{d}r \\ =&\int_0^\infty rh^2(r^2)\int_{-\pi}^{\pi} g^2(\cos\theta)f^2(\sqrt{\epsilon}r,\theta)q(\sqrt{\epsilon}r,\theta)\mathrm{d}\theta\mathrm{d}r \\ =&\int_{\epsilon^{\gamma-\frac{1}{2}}}^\infty+\int_0^{\epsilon^{\gamma-\frac{1}{2}}}\left(rh(r^2)\int_{-\pi}^{\pi} g^2(\cos\theta)f^2(\sqrt{\epsilon}r,\theta)q(\sqrt{\epsilon}r,\theta)\mathrm{d}\theta\right)\mathrm{d}r \\ =:&\,Q_1+Q_2, \end{aligned} \tag{A.1}$$

where $0<\gamma<1/2$. Here

$$\begin{aligned} Q_1=&\int_{\epsilon^{\gamma-\frac{1}{2}}}^\infty rh(r^2)\int_{-\pi}^{\pi} g^2(\cos\theta)f^2(\sqrt{\epsilon}r,\theta)q(\sqrt{\epsilon}r,\theta)\mathrm{d}\theta\mathrm{d}r \\ \le&\, C\exp\big(-\epsilon^{2\gamma-1}\big)=o(\epsilon). \end{aligned}$$

For Q_2, using Taylor expansion

$$f^2(\sqrt{\epsilon}r,\theta)q(\sqrt{\epsilon}r,\theta)=f^2q\big|_{(0,\theta)}+\sqrt{\epsilon}r\left(2fq\frac{\partial f}{\partial r}+f^2\frac{\partial q}{\partial r}\Big|_{(0,\theta)}\right)+\mathcal{O}(\epsilon),$$

we get

$$\begin{aligned} Q_2=&\int_0^{\epsilon^{\gamma-\frac{1}{2}}} rh^2(r^2)\int_{-\pi}^{\pi} g^2(\cos\theta)f^2(\sqrt{\epsilon}r,\theta)q(\sqrt{\epsilon}r,\theta)\mathrm{d}\theta\mathrm{d}r \\ =&\int_0^{\epsilon^{\gamma-\frac{1}{2}}} rh^2(r^2)\int_{-\pi}^{\pi} g^2(\cos\theta)\left(f^2q\big|_{(0,\theta)}+\sqrt{\epsilon}r\left(2fq\frac{\partial f}{\partial r}+f^2\frac{\partial q}{\partial r}\Big|_{(0,\theta)}\right)+\mathcal{O}(\epsilon)\right)\mathrm{d}\theta\mathrm{d}r \\ =&\int_0^\infty rh(r^2)\int_{-\pi}^{\pi} g^2(\cos\theta)\left(f^2q\Big|_{(0,\theta)}+\sqrt{\epsilon}r\left(2fq\frac{\partial f}{\partial r}+f^2\frac{\partial q}{\partial r}\Big|_{(0,\theta)}\right)\right)\mathrm{d}\theta\mathrm{d}r+\mathcal{O}(\epsilon) \\ =&\,\tilde{m}_0f^2(x)q(x)+\sqrt{\epsilon}\tilde{m}_1\mathcal{A}\big(f^2(x)q(x)\big)+\mathcal{O}(\epsilon). \end{aligned} \tag{A.2}$$

For the high-dimensional case, the derivation is similar, so we omit it. □

Appendix B. Proof of Theorem 4.1

Proof. Following [38], we consider using the Chernoff inequality to get an upper bound for $p(n,\alpha)$ with an α-error. We will only estimate the term $P(\mathcal{L}_{\epsilon,n}f-\mathcal{L}_\epsilon f>\alpha)$ since the other part can be made similarly.

Let $\tilde{\alpha}=\sqrt{\epsilon}\alpha$. We have

$$\begin{aligned}p(n,\alpha)&=P\big(\sqrt{\epsilon}(\mathcal{L}_{\epsilon,n}f-\mathcal{L}_\epsilon f)>\tilde{\alpha}\big)\\&=P\left(\frac{\sum_{j=1}^n k_\epsilon(x,s_j)f(s_j)}{\sum_{j=1}^n k_\epsilon(x,s_j)}-\frac{\int k_\epsilon(x,y)f(y)q(y)\mathrm{d}y}{\int k_\epsilon(x,y)q(y)\mathrm{d}y}>\tilde{\alpha}\right).\end{aligned}$$

Since $k_\epsilon(x,s_j)$ is positive, we have

$$p(n,\alpha)=P\left(\sum_{j=1}^n\Big[\mathbb{E}\big(k_\epsilon(x,y)\big)k_\epsilon(x,s_j)f(s_j)-\big(\mathbb{E}\big(k_\epsilon(x,y)f(y)\big)+\tilde{\alpha}\mathbb{E}\big(k_\epsilon(x,y)\big)\big)k_\epsilon(x,s_j)\Big]>0\right),$$

which is equivalent to

$$p(n,\alpha)=P\left(\sum_{j=1}^n Y_j>n\tilde{\alpha}\big(\mathbb{E}\big(k_\epsilon(x,y)\big)\big)^2\right),$$

where

$$\begin{aligned}Y_j:=&\Big[\mathbb{E}\big(k_\epsilon(x,y)\big)k_\epsilon(x,s_j)f(s_j)-\mathbb{E}\big(k_\epsilon(x,y)f(y)\big)k_\epsilon(x,s_j)\Big]\\&+\tilde{\alpha}\mathbb{E}\big(k_\epsilon(x,y)\big)\big(\mathbb{E}\big(k_\epsilon(x,y)\big)-k_\epsilon(x,s_j)\big).\end{aligned}\tag{B.1}$$

We remark that the expectation $\mathbb{E}$ in the above and continued expressions are taken with respect to the variable y or s_j whose probability density function is $q(y)$.

It is easy to find that Y_j are i.i.d random variables with $\mathbb{E}(Y_j)=0$. Next we calculate the variance of Y_j,

$$\mathbb{E}Y_j^2=K_1+K_2+K_3,\tag{B.2}$$

where

$$\begin{aligned}K_1=&\big(\mathbb{E}\big(k_\epsilon(x,y)\big)\big)^2\mathbb{E}\big(k_\epsilon^2(x,y)f^2(y)\big)-2\mathbb{E}\big(k_\epsilon(x,y)\big)\mathbb{E}\big(k_\epsilon(x,y)f(y)\big)\mathbb{E}\big(k_\epsilon^2(x,y)f(y)\big)\\&+\big(\mathbb{E}\big(k_\epsilon(x,y)f(y)\big)\big)^2\mathbb{E}\big(k_\epsilon^2(x,y)\big),\\K_2=&2\tilde{\alpha}\mathbb{E}\big(k_\epsilon(x,y)\big)\big[\mathbb{E}\big(k_\epsilon(x,y)f(y)\big)\mathbb{E}\big(k_\epsilon^2(x,y)\big)-\mathbb{E}\big(k_\epsilon^2(x,y)f(y)\big)\mathbb{E}\big(k_\epsilon(x,y)\big)\big],\\K_3=&\tilde{\alpha}^2\big(\mathbb{E}\big(k_\epsilon(x,y)\big)\big)^2\Big[\mathbb{E}\big(k_\epsilon^2(x,y)\big)-\big(\mathbb{E}\big(k_\epsilon(x,y)\big)\big)^2\Big].\end{aligned}$$

From Lemma A.1, the expectations of $k_\epsilon(x,y)$ and $k_\epsilon(x,y)f(y)$ can be obtained

$$\begin{aligned}\mathbb{E}(k_\epsilon(x,y)) &= \epsilon^{\frac{d}{2}}\left(m_0 q(x)+\sqrt{\epsilon}m_1\hat{v}(x)\cdot\nabla q(x)+\mathcal{O}(\epsilon)\right),\\ \mathbb{E}(k_\epsilon(x,y)f(y)) &= \epsilon^{\frac{d}{2}}\left(m_0 q(x)f(x)+\sqrt{\epsilon}m_1\hat{v}(x)\cdot\nabla\left(q(x)f(x)\right)+\mathcal{O}(\epsilon)\right).\end{aligned} \tag{B.3}$$

From Lemma A.2, the second moments $\mathbb{E}(k_\epsilon^2(x,y)), \mathbb{E}(k_\epsilon^2(x,y)f(y))$ and $\mathbb{E}(k_\epsilon^2(x,y)f^2(y))$ can be obtained

$$\begin{aligned}\mathbb{E}\left(k_\epsilon^2(x,y)\right) &= \epsilon^{\frac{d}{2}}\left(\tilde{m}_0 q(x)+\sqrt{\epsilon}\tilde{m}_1\hat{v}(x)\cdot\nabla q(x)+\mathcal{O}(\epsilon)\right),\\ \mathbb{E}\left(k_\epsilon^2(x,y)f(y)\right) &= \epsilon^{\frac{d}{2}}\left(\tilde{m}_0 q(x)f(x)+\sqrt{\epsilon}\tilde{m}_1\hat{v}(x)\cdot\nabla\left(q(x)f(x)\right)+\mathcal{O}(\epsilon)\right),\\ \mathbb{E}\left(k_\epsilon^2(x,y)f^2(y)\right) &= \epsilon^{\frac{d}{2}}\left(\tilde{m}_0 q(x)f^2(x)+\sqrt{\epsilon}\tilde{m}_1\hat{v}(x)\cdot\nabla\left(q(x)f^2(x)\right)+\mathcal{O}(\epsilon)\right).\end{aligned} \tag{B.4}$$

Substituting (B.3), (B.4) into (B.2), we get

$$\begin{aligned}\mathbb{E}Y_j^2 = {}& 2\tilde{\alpha}\mathbb{E}\left(k_\epsilon(x,y)\right)\left[\mathbb{E}\left(k_\epsilon(x,y)f(y)\right)\mathbb{E}\left(k_\epsilon^2(x,y)\right)-\mathbb{E}\left(k_\epsilon^2(x,y)f(y)\right)\mathbb{E}\left(k_\epsilon(x,y)\right)\right]\\ &+\tilde{\alpha}^2\left(\mathbb{E}\left(k_\epsilon(x,y)\right)\right)^2\left[\mathbb{E}\left(k_\epsilon^2(x,y)\right)-\left(\mathbb{E}\left(k_\epsilon(x,y)\right)\right)^2\right]+C\epsilon^{\frac{3d+2}{2}},\end{aligned} \tag{B.5}$$

where C is a constant.

Note that our interested regime is $\tilde{\alpha}\lesssim\mathcal{O}(\epsilon)$ since we are estimating an $\mathcal{O}(\sqrt{\epsilon})$ quantity, namely $\sqrt{\epsilon}\mathcal{L}f(x)$, so an error $\tilde{\alpha}$ larger than the estimated quantity is meaningless. For the $\tilde{\alpha}$ term in (B.5), straightforward calculations based on (B.3)-(B.4) show

$$\mathbb{E}\left(k_\epsilon(x,y)\right)\left[\mathbb{E}\left(k_\epsilon(x,y)f(y)\right)\mathbb{E}\left(k_\epsilon^2(x,y)\right)-\mathbb{E}\left(k_\epsilon^2(x,y)f(y)\right)\mathbb{E}\left(k_\epsilon(x,y)\right)\right]=\mathcal{O}\left(\epsilon^{\frac{3d+1}{2}}\right).$$

Thus, the $\tilde{\alpha}$ and $\tilde{\alpha}^2$ terms of the variance (B.5) are negligible, and we obtain

$$\mathbb{E}Y_j^2 = C_1\epsilon^{\frac{3d+2}{2}},$$

where C_1 is a constant. By the Chernoff inequality, we obtain

$$p(n,\alpha)\le 2\exp\left(-\frac{n\tilde{\alpha}^2\epsilon^{\frac{d}{2}}}{C_1\epsilon}\right)=2\exp\left(-\frac{n\alpha^2\epsilon^{\frac{d}{2}}}{C_1}\right). \tag{B.6}$$

The inequality (B.6) means that the correct magnitude of α should be made such that

$$n\alpha^2\epsilon^{\frac{d}{2}}=\mathcal{O}(1), \tag{B.7}$$

i.e. $\alpha\sim\mathcal{O}(n^{-1/2}\epsilon^{-d/4})$. □

Appendix C. Pseudocode for the algorithms

In this appendix, we present the pseudocode of the algorithms introduced in this paper.

Algorithm 1 Time Rescaling (Proposal 2.1).

Input: Latent time matrix $T\in\mathbb{R}^{n\times d}$.

1: Compute $W=\mathrm{diag}(w_1,\cdots,w_d)$ in which $w_g=1/\|t_{\bullet g}\|$, $g=1,\ldots,d$.
2: Get eigenvalues and eigenvectors $[\lambda,V]=\mathrm{eig}(W^\top T^\top TW)$.
3: $\lambda_1=\lambda[0]$ be the maximum eigenvalue.
4: $v_1=V[:,0]$ be the eigenvector corresponding to the maximum eigenvalue.
5: $x=d\lambda_1^{-1/2}Wv_1,\beta=1/x$.
6: $t=TWv_1/\|TWv_1\|$.

Output: t,β,x.

Algorithm 2 Time Rescaling (Proposal 2.2).

Input: Latent time matrix $T\in\mathbb{R}^{n\times d}$.

1: Get eigenvalues and eigenvectors $[\lambda,V]=\mathrm{eig}(T^\top T)$.
2: $v_1=V[:,0]$ be the eigenvector corresponding to the maximum eigenvalue.
3: $\beta=v_1,x=1/\beta$.
4: $t=Tv_1$.

Output: t,β,x.

Algorithm 3 SEM Algorithm for Parameter Uncertainty Quantification.

Input: θ^*, i.e. $\hat{\theta}_n$ and the sequence $\theta^{(k)},k=1,\ldots,n$.

1: **for** i=1:g **do**
2: Let r_{ij} be the (i,j)-th element of Jacobian J_M.
3: $\theta^{(k)}(i)=(\theta_1^*,\cdots,\theta_{i-1}^*,\theta_i^{(k)},\theta_{i+1}^*,\cdots,\theta_d^*)$ which means that we only iterate on the i-th component $\theta_i^{(k)}$, while the other components are fixed.
4: For given $\theta^{(k)}(i)$, iterate the one-step EM algorithm to obtain $\tilde{\theta}^{(k+1)}(i)$.
5: To get difference quotient

$$r_{ij}^{(k)}=\frac{\tilde{\theta}_j^{(k+1)}(i)-\theta_j^*}{\theta_i^{(k)}-\theta_i^*},\quad j=1,\ldots,g.$$

6: **end for**
7: Obtain r_{ij} using the least squares method when the sequence $r_{ij}^{(k^*)},r_{ij}^{(k^*+1)},\cdots$ is stable for some k^*.
8: Linearize $L(\theta\,|\,x_{obs},t)$ and get $\hat{I}_{oc}$

$$\hat{I}_{oc}=\hat{I}_{oc}^{-1}\big(\theta^*\,|\,S(x_{obs},t)\big),$$

where $S(x_{obs},t)$ is obtained at the last E step.

Output: $\hat{I}_{oc},J_M$.

Algorithm 4 First Hitting Time to Set A.

Input: Transition probability matrix $P\in\mathbb{R}^{n\times n}$, stop criteria N, cell set A.

1: Construct matrix $Q\in\mathbb{R}^{n\times n}, Q=(q_{ij}):=(p_{ij})_{i,j\in A^c}$.
2: Initialize $K_0=[0,0,\cdots,0]^\top\in\mathbb{R}^n$.
3: **for** i=1:N **do**
4: $\quad K_n=1+QK_{n-1}$.
5: **end for**

Output: K_N.

Algorithm 5 First Hitting Time to Set A (with taboo set H).

Input: Transition probability matrix $P\in\mathbb{R}^{n\times n}$, stop criteria N, cell set A, taboo set H.

1: Construct matrix $Q\in\mathbb{R}^{(n-H)\times(n-H)}, Q=(q_{ij}):=(p_{ij})_{i,j\in A^c\cup H^c}$.
2: Initialize $K_0=[0,0,\cdots,0]^\top\in\mathbb{R}^{n-|H|}$.
3: **for** i=1:N **do**
4: $\quad K_n=1+QK_{n-1}$.
5: **end for**

Output: K_N.

Acknowledgments

We thank Dr. Yichong Wu for the supportive starting work in this project.

T. Li, Y. Wang and G. Yang acknowledge the support from the MSTC (Grant No. 2021YFA1003301), and from the NSFC (Grant Nos. 11825102, 12288101). P. Zhou is supported by Start-up Grants of the Peking University (Grant No. 7101303365).

References

[1] L. Atta, A. Sahoo, and J. Fan, *Veloviz: RNA velocity-informed embeddings for visualizing cellular trajectories*, Bioinformatics, 38(2):391–396, 2022.
[2] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*, Springer, 1999.
[3] V. Bergen, M. Lange, S. Peidli, F. A. Wolf, and F. J. Theis, *Generalizing RNA velocity to transient cell states through dynamical modeling*, Nat. Biotechnol., 38(12):1408–1414, 2020.
[4] F. Bocci, P. Zhou, and Q. Nie, *SpliceJAC: Transition genes and state-specific gene regulation from single-cell transcriptome data*, Mol. Syst. Biol., 18(11):e11176, 2022.
[5] R. T. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, *Neural ordinary differential equations*, Adv. Neural Inf. Process. Syst., 31:6572–6583, 2018.
[6] Z. Chen, W. C. King, A. Hwang, M. Gerstein, and J. Zhang, *DeepVelo: Single-cell transcriptomic deep velocity field learning with neural ordinary differential equations*, Sci. Adv., 8:eabq3745, 2022.
[7] J. Dong, P. Zhou, Y. Wu, Y. Chen, H. Xie, Y. Gao, J. Lu, J. Yang, X. Zhang, L. Wen, T. Li, and F. Tang, *Integrating single-cell datasets with ambiguous batch information by incorporating molecular network features*, Brief. Bioinformatics, 23(1):bbab366, 2022.

[8] N. Eling, M. D. Morgan, and J. C. Marioni, *Challenges in measuring and understanding biological noise*, Nat. Rev. Genet., 20(9):536–548, 2019.
[9] S. Farrell, M. Mani, and S. Goyal, *Inferring single-cell transcriptomic dynamics with structured dynamical representations of RNA velocity*, Bulletin of the American Physical Society, Vol. 68(3):N10.00002, 2023.
[10] M. Gao, C. Qiao, and Y. Huang, *UniTVelo: Temporally unified RNA velocity reinforces single-cell trajectory inference*, Nat. Commun., 13(1):6586, 2022.
[11] A. Gayoso, P. Weiler, M. Lotfollahi, D. Klein, J. Hong, A. M. Streets, F. J. Theis, and N. Yosef, *Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells*, Nat. Methods, 2023, doi: 10.1038/s41592-023-01994-w.
[12] G. Gorin, M. Fang, T. Chari, and L. Pachter, *RNA velocity unraveled*, PLoS Comput. Biol., 18(9):e1010492, 2022.
[13] G. Gorin and L. Pachter, *Modeling bursty transcription and splicing with the chemical master equation*, Biophys. J., 121(6):1056–1069, 2022.
[14] G. Gorin, J. J. Vastola, M. Fang, and L. Pachter, *Interpretable and tractable models of transcriptional noise for the rational design of single-molecule quantification experiments*, Nat. Commun., 13:7620, 2022.
[15] Y. Gu, D. Blaauw, and J. D. Welch, *Bayesian inference of RNA velocity from multi-lineage single-cell data*, bioRxiv, 2022, doi: 10.1101/2022.07.08.499381v1.
[16] L. Haghverdi, M. Büttner, F. A. Wolf, F. Buettner, and F. J. Theis, *Diffusion pseudotime robustly reconstructs lineage branching*, Nat. Methods, 13(10):845–848, 2016.
[17] T. Hashimoto, Y. Sun, and T. Jaakkola, *From random walks to distances on unweighted graphs*, Adv. Neural Inf. Process. Syst., 28:3429–3437, 2015.
[18] A. Hilfinger and J. Paulsson, *Separating intrinsic from extrinsic fluctuations in dynamic biological systems*, Proc. Natl. Acad. Sci. USA, 108(29):12167–12172, 2011.
[19] R. Horn and C. Johnson, *Matrix Analysis*, Cambridge University Press, 1985.
[20] J. Kemeny and J. Snell, *Finite Markov Chains*, Springer, 1976.
[21] G. La Manno et al., *RNA velocity of single cells*, Nature, 560(7719):494–498, 2018.
[22] S. S. Lafon, *Diffusion Maps and Geometric Harmonics*, Yale University, 2004.
[23] M. Lange et al., *CellRank for directed single-cell fate mapping*, Nat. Methods, 19(2):159–170, 2022.
[24] E. L. Lehmann and G. Casella, *Theory of Point Estimation*, Springer, 1998.
[25] C. Li, M. C. Virgilio, K. L. Collins, and J. D. Welch, *Multi-omic single-cell velocity models epigenome-transcriptome interactions and improves cell fate prediction*, Nat. Biotechnol., 41(3):387–398, 2023.
[26] T. Li, J. Shi, Y. Wu, and P. Zhou, *On the mathematics of RNA velocity I: Theoretical analysis*, CSIAM Trans. Appl. Math, 2(1):1–55, 2021.
[27] R. Liu, A. O. Pisco, E. Braun, S. Linnarsson, and J. Zou, *Dynamical systems model of RNA velocity improves inference of single-cell trajectory, pseudo-time and gene regulation*, J. Mol. Biol., 434:167606, 2022.
[28] J. Ma, J. Zhao, J. Tian, X. Bai, and Z. Tu, *Regularized vector field learning with sparse approximation for mismatch removal*, Pattern Recognit., 46(12):3519–3532, 2013.
[29] V. Marot-Lassauzaie, B. J. Bouman, F. D. Donaghy, Y. Demerdash, M. A. G. Essers, and L. Haghverdi, *Towards reliable quantification of cell state velocities*, PLoS Comput. Biol., 18(9): e1010031, 2022.
[30] L. McInnes, J. Healy, N. Saul, and L. Großberger, *UMAP: Uniform manifold approximation and projection*, J. Open Source Softw., 3(29):861, 2018.

[31] L. Meng and J. C. Spall, *Efficient computation of the Fisher information matrix in the EM algorithm*, in: 51st Annual Conference on Information Sciences and Systems (CISS), IEEE, 2017, 1–6.
[32] X.-L. Meng and D. B. Rubin, *Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm*, J. Am. Stat. Assoc., 86(416):899–909, 1991.
[33] J. R. Norris, *Markov Chains*, Cambridge University Press, 1997.
[34] C. Qiao and Y. Huang, *Representation learning of RNA velocity reveals robust cell transitions*, Proc. Natl. Acad. Sci. USA, 118(49):e2105859118, 2021.
[35] Q. Qin, E. Bingham, G. La Manno, D. M. Langenau, and L. Pinello, *Pyro-Velocity: Probabilistic RNA velocity inference from single-cell data*, bioRxiv, 2022, doi: 10.1101/2022.09.12.507691.
[36] X. Qiu et al., *Mapping transcriptomic vector fields of single cells*, Cell, 185(4):690–711.e45, 2022.
[37] M. Setty, V. Kiseliovas, J. Levine, A. Gayoso, L. Mazutis, and D. Pe'er, *Characterization of cell fate probabilities in single-cell data with palantir*, Nat. Biotechnol., 37:451–460, 2019.
[38] A. Singer, *From graph to manifold Laplacian: The convergence rate*, Appl. Comput. Harmon. Anal., 21(1):128–134, 2006.
[39] J. C. Spall, *Monte Carlo computation of the Fisher information matrix in nonstandard settings*, J. Comput. Graph. Stat., 14(4):889–909, 2005.
[40] S. Spigler, M. Geiger, and M. Wyart, *Asymptotic learning curves of kernel methods: Empirical data versus teacher-student paradigm*, J. Stat. Mech. Theory Exp., 2020:124001, 2020.
[41] S. V. Stassen, G. G. Yip, K. K. Wong, J. W. Ho, and K. K. Tsia, *Generalized and scalable trajectory inference in single-cell omics data with VIA*, Nat. Commun., 12(1):5528, 2021.
[42] F. Tang et al., *mRNA-Seq whole-transcriptome analysis of a single cell*, Nat. Methods, 6:377–382, 2009.
[43] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, *Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions*, Nat. Biotechnol., 32:381, 2014.
[44] U. Von Luxburg, A. Radl, and M. Hein, *Hitting and commute times in large random neighborhood graphs*, J. Mach. Learn. Res., 15(52):1751–1798, 2014.
[45] Y. Wang and T. Li, *Inferring the RNA velocity with stochastic models*, Peking University, preprint, 2023.
[46] Y. Wang and Z. Wang, *Inference on the structure of gene regulatory networks*, J. Theor. Biol., 539:111055, 2022.
[47] C. Weinreb, S. Wolock, B. K. Tusi, M. Socolovsky, and A. M. Klein, *Fundamental limits on dynamic inference from single-cell snapshots*, Proc. Natl. Acad. Sci. USA, 115(10):E2467–E2476, 2018.
[48] F. A. Wolf, F. K. Hamey, M. Plass, J. Solana, J. S. Dahlin, B. Göttgens, N. Rajewsky, L. Simon, and F. J. Theis, *PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells*, Genome Biol., 20:1–9, 2019.
[49] S. Y. Zhang and M. P. H. Stumpf, *Learning cell-specific networks from dynamical single cell data*, bioRxiv, 2023, doi: 10.1101/2023.01.08.523176v2.
[50] Z. Zhang and X. Zhang, *Inference of high-resolution trajectories in single-cell RNA-seq data by using RNA velocity*, Cell Rep. Methods, 1(6):100095, 2021.
[51] P. Zhou, X. Gao, X. Li, L. Li, C. Niu, Q. Ouyang, H. Lou, T. Li, and F. Li, *Stochasticity triggers activation of the S-phase checkpoint pathway in budding yeast*, Phys. Rev. X, 11:011004, 2021.
[52] P. Zhou, S. Wang, T. Li, and Q. Nie, *Dissecting transition cells from single-cell transcriptome data through multiscale stochastic dynamics*, Nat. Commun., 12:5609, 2021.