Title

Application of Quantile Generalized Additive Model in Differential Expressed Gene Inference for Single-cell RNA Sequencing Data

Author

Liu, Tianyang

Publication Date

2022

UNIVERSITY OF CALIFORNIA

Los Angeles

Application of Quantile Generalized Additive Model
in Differential Expressed Gene Inference
for Single-cell RNA Sequencing Data

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Applied Statistics

by

Tianyang Liu

2022

ABSTRACT OF THE THESIS

Application of Quantile Generalized Additive Model
in Differential Expressed Gene Inference
for Single-cell RNA Sequencing Data

by

Tianyang Liu
Master of Applied Statistics
University of California, Los Angeles, 2022
Professor Jingyi Li, Chair

This thesis focuses on applying the quantile generalized additive model (QGAM) to detect differentially expressed (DE) genes in the single-cell RNA sequencing data. Most of the existing DE gene inference methods are developed based on Generalized Additive Model (GAM). At the same time, GAM can be sensitive to outliers during the model fitting process. Such sensitivity impacts the accuracy of the DE gene detection for the dataset with outliers. We want to use QGAM's robustness to outliers to improve the DE gene detection accuracy. We compared the performance of the QGAM-based DE gene inference method (qgamDE) with two state-of-the-art GAM-based methods, PseudotimeDE, and tradeSeq, by applying them to the simulated data. In conclusion, the performance of qgamDE and PseudotimeDE is better and more stable compared to the performance of tradeSeq. As a result, we added qgamDE to the PseudotimeDE R package as new functionality.

The thesis of Tianyang Liu is approved.

Yingnian Wu

Frederic R. Paik Schoenberg

Jingyi Li, Committee Chair

University of California, Los Angeles

2022

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to thank my thesis committee chair, Professor Jingyi Li, for her suggestions in this thesis. I also want to thank Mr. Dongyuan Song for introducing this topic to me. I finally would like to thank my thesis committee members for all of their help on my thesis.

CHAPTER 1

Introduction

Single-cell RNA sequencing (scRNA-seq) has recently developed rapidly, promoting research in biologies such as transcriptomic heterogeneity dissertation and cell type or state detection [3, 6]. Since scRNA-seq records the continuous cell transition process, it facilitates the research studies that involve cellular process exploration. However, the data analysis for scRNA-seq data can be a challenge. Pseudotime, which was first proposed by Trapnell et al. in 2004 [13], is a sequence of values to represent the stages associated with gene development. Some pseudotime/trajectory inference methods have been developed; for example, TSCAN [5], Slingshot [12], and Monocle [13]. Those methods arrange cells into lineages with pseudotime-based order, then the pattern of a dynamic process experienced by cells can be determined.

The DE gene identification is a downstream analysis of trajectory inference. The workflow of this process can be described as follows: once we have a scRNA-seq dataset, we can infer a pseudotime vector and align the pseudotime with genes to show their development at different stages. When we have the gene expression and its corresponding pseudotime, we can fit a regression model by setting pseudotime as the predictor and the gene expression as the response variable. According to the fitted model, if the gene expression is related to the pseudotime, this gene can be regarded as a DE gene. Otherwise, the gene is a non-DE gene.

Currently, some DE gene inference methods have been developed. PseudotimeDE [11] and tradeSeq [14] are two state-of-the-art methods. They handle the DE gene inference task by applying a generalized additive model (GAM) to find the relationship between the gene expression in cells and the corresponding pseudotime. However, GAM can be sensitive to outliers because it fits the mean of the response variable. If there are outliers in the dataset, the GAM-based inference methods might not precisely find the relationship between the gene expression and the corresponding pseudotime. Here we would like to introduce a technique, quantile generalized additive model (QGAM), developed by Fasiolo et al. [1]. Since QGAM fits the specific quantile of the response variable, the outlier does not impact the fitted result. Based on this idea, we adopted QGAM to infer the relationship between the gene expression and corresponding pseudotime for the scRNA-seq data.

In this thesis, we propose the QGAM-based DE gene inference method (qgamDE) to do the DE gene inference task. To show the outlier robustness of our approach in DE gene detection, we simulated two sets of data for the comparison experiment. One data set contains more outliers, while another data set contains fewer outliers. This project focuses on one type of outliers, doublets, which are artifactual libraries generated from two cells in scRNA-seq experiments. A doublet is a single cell generated by mistakenly adding two cells together. According to Zheng et al. [17], the doublets usually appears because of errors in cell sorting or capture. In single-cell level research, doublets are undesirable. They can interfere with DE gene analysis, produce spurious cell clusters, and obscure the inference of cell developmental trajectories [8, 15]. The existing experiment result shows that qgamDE can achieve relatively high AUC and power in DE gene detection. Furthermore, its performance does not obviously deteriorate as the number of doublets increases.

The structure of this thesis is as follows: Chapter 2 reviews the critical concepts of quantile regression, GAM, QGAM, and three DE gene inference methods. Chapter 3 shows the data simulation process and the visualization of the simulated datasets. Chapter 4 compares the performance of tradeSeq, PseudotimeDE, and qgamDE in the DE gene inference task.

Chapter 5 discusses an issue with qgamDE and the current state of the qgamDE application. Chapter 6 concludes the thesis and discusses possible improvement of qgamDE.

CHAPTER 2

Methodology

2.1 Quantile Regression

In this thesis, we only consider the scenario that there are only one covariate vector $\boldsymbol{X} \in \mathbb{R}^{n\times 1}$ and one response variable vector $\boldsymbol{Y} \in \mathbb{R}^{n\times 1}$. Before we jump into the actual quantile generalized additive model, we start with the basic concepts of quantile regression. The main difference between the linear regression and quantile regression is that the linear regression model predicts the conditional mean $E[Y|X]$. In contrast, quantile regression predicts the τ-th conditional quantile $q(X) = \inf\{Y : F(Y \mid X) \geq \tau\}$, where $F(Y \mid X)$ is the conditional c.d.f. of Y [1]. The traditional quantile regression model is set as

$$q(X_i) = X_i\beta,$$

where β is the quantile regression coefficient. According to Koenker et al. [7], the coefficient estimator $\hat{\beta}$ for the τ-th quantile can be obtained by L_1 weighted loss function

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \rho_\tau \left\{Y_i - X_i\beta\right\},$$

where X_i is the i-th element the vector $\boldsymbol{X}$, Y_i is the i-th element the vector $\boldsymbol{Y}$, and ρ_τ is a loss function

$$\rho_\tau(e) = e(\tau - \mathbb{I}(e < 0)) = \begin{cases} e(\tau - 1), e < 0, \\ e\tau, e \geq 0. \end{cases}$$

2.2 Generalized Additive Model

After introducing the regular quantile regression model, let us move to the generalized additive model (GAM). GAM, a non-parametric regression model, was developed by Hastie et al. [4]. This model is formed by summing up many functions of predictors. The model itself tries to fit the mean of the response variable, where the response variable is from an exponential family. The actual model structure can be described as

$$g\left(\mu_i\right) = f\left(X_i\right),$$

where μ_i is the expectation of the response variable $E(Y_i)$, f is the smooth function of predictor, and g is the link function. GAM is classified as a non-parametric method because it allows describing the relationship between the response variable and the predictors by the smooth function instead of the specific parameters.

2.3 Quantile Generalized Additive Model

Now, with the concepts of quantile regression and generalized additive model, we are ready for the quantile generalized additive model (QGAM). In QGAM, Fasiolo et al. [1] assume that the τ-th quantile of the response variable depends on the functions of predictor

$$q(X_i) = f(X_i).$$

The marginal smooth effect additive term is

$$f(X_i) = \sum_{k=1}^{r} \beta_k b_k\left(X_i\right),$$

where β_k are the unknown coefficients and $b_k(\cdot)$ are known spline basis functions [1]. To avoid over-fitting, the complex level of QGAM needs to be controlled. Then the penalized loss is defined as

$$V(\beta, \gamma, \sigma) = \sum_{i=1}^{n} \frac{1}{\sigma} \rho_\tau \left\{ Y_i - f(X_i) \right\} + \frac{1}{2} \gamma \boldsymbol{\beta}^T \mathbf{S} \boldsymbol{\beta},$$

where $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_r]^T$, γ is a positive smoothing parameter, $\frac{1}{\sigma} > 0$ is the learning rate which is used to determine the relative weight of the loss and the penalty, $\mathbf{S}$ is a positive semi-definite matrix and it is used to penalize the wiggliness of the corresponding effect f, the function $\rho_\tau(e)$ is the same loss function which we have introduced in the quantile regression section [1].

2.4 PseudotimeDE

PseudotimeDE is a pseudotime-based differential expression method developed by Song and Li in 2021 [11]. It considers the random nature of inferred pseudotime by applying subsampling to estimate pseudotime inference uncertainty and propagating the uncertainty to its statistical test for DE gene identification [11].

In PseudotimeDE, the baseline model that describes the relationship between the gene expression and the pseudotime is the negative binomial–generalized additive model (NB-GAM). For gene j ($j \in \{1, \ldots, m\}$), its expression Y_{ij} in cell i and the pseudotime T_i of cell i ($i \in \{1, \ldots, n\}$) are assumed to follow

$$\begin{cases} Y_{ij} \sim \mathrm{NB}\left(\mu_{ij}, \phi_j\right), \\ \log\left(\mu_{ij}\right) = \beta_{j0} + f_j\left(T_i\right), \end{cases}$$

where $\mathrm{NB}(\mu_{ij}, \phi_j)$ is the negative binomial distribution with mean μ_{ij} and dispersion ϕ_j, and $f_j\left(T_i\right) = \sum_{k=1}^{r} b_k\left(T_i\right) \beta_{jk}$ is a cubic spline function [11].

To test if gene j is differentially expressed along cell pseudotime, PseudotimeDE defines the null and alternative hypotheses as $H_0 : f_j(\cdot) = 0$ vs. $H_1 : f_j(\cdot) \neq 0$ [11]. According to Wood [16], the test statistic is set as

$$S_j = \hat{\boldsymbol{f}}_j^{\mathrm{T}} \hat{\mathbf{V}}_{f_j}^{r-} \hat{\boldsymbol{f}}_j,$$

where $\hat{\mathbf{V}}_{f_j}^{r-}$ is a rank-r pseudo-inverse of $\hat{\mathbf{V}}_{f_j}$ and $\hat{\mathbf{V}}_{f_j}$ is the estimated covariance matrix of $\hat{\boldsymbol{f}}_j$, $\hat{\boldsymbol{f}}_j$ is the estimates of a vector of $f_j(\cdot)$ on each T_i, $[f_j(T_1), \ldots, f_j(T_n)]^T$.

In PseudotimeDE, 80% cells (rows) are subsampled from gene expression count matrix $\mathbf{Y}$ for B times to estimate the uncertainty of pseudotime. Suppose the cells in the original data have been classified into some groups. In that case, PseudotimeDE will first subsample 80% cells from each group and then combine all the within-group subsamples into one subsample [11]. For each subsample $\mathbf{Y}^b = [Y_{ij}]^b$, an $n' \times m$ matrix where $n' = \lfloor 0.8n \rfloor$, PseudotimeDE will apply the pseudotime inference method to the subsampled data with the same input parameters applied to the original dataset. As a result, B subsample-based realizations of pseudotime $\mathbf{T} : \left\{\boldsymbol{T}^1, \cdots, \boldsymbol{T}^b, \cdots, \boldsymbol{T}^B\right\}$ are generated [11]. One subsampled pseudotime set can be described as $\boldsymbol{T}^b = [T_1^b, \ldots, T_{n'}^b]^T$. Note that the pseudotime inference process for each subsample comes before the permutation process for each subsample. In PseudotimeDE, the subsampled n' pseudotime T^b could be the n' values that are not in the original pseudotime, then the uncertainty in pseudotime inference can be reflected in $\boldsymbol{T}^b$ [11].

PseudotimeDE does the permutation as follows: First, PseudotimeDE randomly permutes each subsampled pseudotime $\boldsymbol{T}^b = [T_1^b, \ldots, T_{n'}^b]^T$ into $\boldsymbol{T}^{*b} = [T_1^{*b}, \ldots, T_{n'}^{*b}]^T$ [11]. Second, PseudotimeDE fits the previous model to $[Y_{1j}^b, \ldots, Y_{n'j}^b]^T$ and $\boldsymbol{T}^{*b}$ and calculates the test statistic S_j's value as s_j^b [11], Third, PseudotimeDE repeats previous two steps for $b \in \{1, \ldots, B\}$ and stores $\left\{s_j^1, \ldots, s_j^B\right\}$ as the null values of the test statistic S_j.

PseudotimeDE estimates the null distribution of S_j in two ways: Empirical estimate and Parametric estimate. PseudotimeDE calculates the p-value for gene j based on the estimated null distribution in either way and the observed test statistic value s_j [11]. Empirical estimate: The empirical distribution of $\left\{s_j^1, \ldots, s_j^B\right\}$ is used as the estimated null distribution [11]. According to Phipson et al. [9], the p-value of gene j is calculated as

$$p_j^{emp} = \frac{\sum_{b=1}^{B} \mathbb{I}\left(s_j^b \geq s_j\right) + 1}{B+1},$$

where $\mathbb{I}(\cdot)$ is the indicator function. Parametric estimate: PseudotimeDE fits a parametric distribution to $\left\{s_j^1, \ldots, s_j^B\right\}$ and the fitted distribution is used as the estimated null distribution [11]. Two parametric distributions, a gamma distribution $\Gamma(\alpha, \beta)$ with $\alpha, \beta > 0$ and a two-component gamma mixture model $\gamma\Gamma\left(\alpha_1, \beta_1\right) + (1-\gamma)\Gamma\left(\alpha_2, \beta_2\right)$ with $0 < \gamma < 1$ and $\alpha_1, \beta_1, \alpha_2, \beta_2 > 0$, are considered by PseudotimeDE. After fitting both distributions to $\left\{s_j^1, \ldots, s_j^B\right\}$ using the maximum likelihood estimation, PseudotimeDE selects one of the fitted distributions by using the likelihood ratio test with 3 degrees of freedom [11]. If the likelihood ratio test p-value is greater than 0.01, PseudotimeDE uses the fitted gamma distribution as the parametric estimate of the null distribution of S_j; otherwise, PseudotimeDE uses the fitted two-component gamma mixture model. PseudotimeDE calculates the p-value of gene j as

$$p_j^{param} = 1 - \hat{F}_j\left(s_j\right),$$

where $\hat{F}_j(\cdot)$ is the cumulative distribution function of the parametrically estimated null distribution.

2.5 tradeSeq

tradeSeq, a method for trajectory-based differential expression analysis for sequencing data, was developed by Van den Berge et al. [14]. tradeSeq detects DE genes based on pseudotime inference. Moreover, tradeSeq also uses GAM to infer the relationship between gene expression and pseudotime.

In the process of DE gene detection, the read counts Y_{ij}, for a given gene $j \in \{1, \ldots, m\}$ across cells $i \in \{1, \ldots, n\}$ are modeled using a NB-GAM with cell and gene-specific means μ_{ij} and gene-specific dispersion parameters ϕ_j [14]

$$\begin{cases} Y_{ij} \sim \mathrm{NB}\left(\mu_{ij}, \phi_j\right), \\ \log\left(\mu_{ij}\right) = \eta_{ij}, \\ \eta_{ij} = \sum_{l=1}^{L} s_{lj}\left(T_{li}\right) Z_{li} + \mathbf{U}_i \boldsymbol{\alpha}_j + \log\left(N_i\right). \end{cases}$$

The gene-wise additive predictor η_{ij} is composed of lineage-specific smoothing splines s_{lj}, and s_{lj} are functions of pseudotime T_{li}, for lineages $l \in \{1, \ldots, L\}$ [14]. The binary matrix $\mathbf{Z} = (Z_{li} \in \{0, 1\} : l \in \{1, \ldots, L\}, i \in \{1, \ldots, n\})$ assigns every cell to a specific lineage based on some weights [14]. $\mathbf{U}$ is an $n \times p$ matrix, where $\mathbf{U}_i$ represents the i-th cell. $\boldsymbol{\alpha_j}$ is the regression parameter with dimension of $p \times 1$. In general, the matrix $\mathbf{U}$ contains the effects of p known cell-level covariates(e.g., batch, age, or gender). N_i are the cell-specific offsets which account for the differences in sequencing depth or the capture efficiency between cells [14].

s_{lj} is the smoothing spline for a given gene j and lineage l. It can be represented as a linear combination of r cubic basis functions

$$s_{lj}(t) = \sum_{k=1}^{r} b_k(t) \beta_{ljk},$$

where the cubic basis functions $b_k(t)$ are the same for all genes and lineages [14].

The null hypothesis for DE gene testing is set as H_0: $\mathbf{C}^T \boldsymbol{\beta}_j = 0$ ($\boldsymbol{\beta}_j$ is the concatenation of the Lr-dimensional column vectors β_{lj} of lineage-specific smoother coefficients). The test statistics is

$$S_j = \hat{\boldsymbol{\beta}}_j^T \mathbf{C} \left(\mathbf{C}^T \hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_j} \mathbf{C} \right)^{-1} \mathbf{C}^T \hat{\boldsymbol{\beta}}_j,$$

where $\hat{\boldsymbol{\beta}}_j$ is an estimator of $\boldsymbol{\beta}_j$, $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}_j}$ is an estimator of the covariance matrix $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}_j}$, and $\mathbf{C}$ is an $(Lr) \times C$ matrix representing the C contrasts of interest for the DE test [14].

2.6 qgamDE

We applied two techniques, QGAM and the pseudotime inference method, to make the DE gene inference. The main difference between our DE gene detection method and the previous two methods is that instead of using GAM to find the relationship between gene expression and pseudotime, we use QGAM to determine the relationship.

For a $n \times m$ gene expression count matrix $\mathbf{Y}$ with rows as cells and columns as genes, we use trajectory inference method to obtain the corresponding pseudotime $\boldsymbol{T} = [T_1, \ldots, T_n]^T$. Once we have the count matrix and the pseudotime, we can use QGAM as the baseline model to find the relationship. For an expression Y_{ij} in τ-th quantile of gene j

$$\log(q_j(T_i)) = \beta_{j0} + f_j(T_i),$$

where $f_j(\mathrm{T_i}) = \sum_{k=1}^{r} \beta_{jk} b_{jk}(T_i)$.

For the DE gene testing, the null and alternative hypotheses are set as $H_0 : f_j(\cdot) = 0$ v.s $H_0 : f_j(\cdot) \neq 0$. The test statistics is set as

$$S_j = \hat{\boldsymbol{f}}_j^T \hat{\mathbf{V}}_{f_j}^{r-} \hat{\boldsymbol{f}}_j,$$

where $\hat{\mathbf{V}}_{f_j}^{r-}$ and $\hat{\boldsymbol{f}}_j$ have been described in PseudotimeDE section [11].

CHAPTER 3

Experiment

3.1 Data Simulation

To evaluate the performance of qgamDE in DE gene detection, we simulated the scRNA-seq data with two cell trajectories: single-lineage and bifurcation.

3.1.1 Single-lineage Dataset

We simulated 500 cells and 1000 genes for the single-lineage data. We generated 500 values from a uniform distribution with a range of $[0, 1]$ as the default time values t for the cells. Then we created a function to simulate the relationship between the time values and the log mean of the gene expression. The function we used is

$$\log(\mu) = a + b \cdot \cos(t) + c \cdot t^2 + d \cdot \sin(t^3) + 1,$$

where a, b, c, d are random variables with $a \sim \text{Uniform}(0, 1)$, $b \sim \text{Uniform}(2, 3)$, $c \sim \text{Uniform}(-2, 2)$, $d \sim \text{Uniform}(-4, 4)$. The ultimate goal of setting the relation function like this is that we do not want all the genes to have the same cell trajectory.

We generated 1000 genes with different trajectories from the negative binomial distribution with the simulated mean of the gene expression. Then we generated another 500 genes with the constant gene expression mean value as the non-DE genes. After that, we randomly selected half of the previous 1000 genes and replaced them with the prepared non-DE genes. In the end, we obtained a dataset with half of the genes being DE genes (with trajectory) and half of the genes being non-DE genes (with no trajectory)

Once we obtaind the simulated single-lineage dataset, we used the trajectory inference method, Slingshot, to obtain the pseudotime for the genes. During the pseudotime inference process, since the cell cluster input is required for $Slingshot()$, we have to assign cell clusters prior to the pseudotime inference. The cell clusters will not bother the inference result for the single-lineage data. We assigned the first half of the cells to cluster 1 and the remaining half to cluster 2. Here are the PCA and UMAP plots of the simulated single-lineage data:

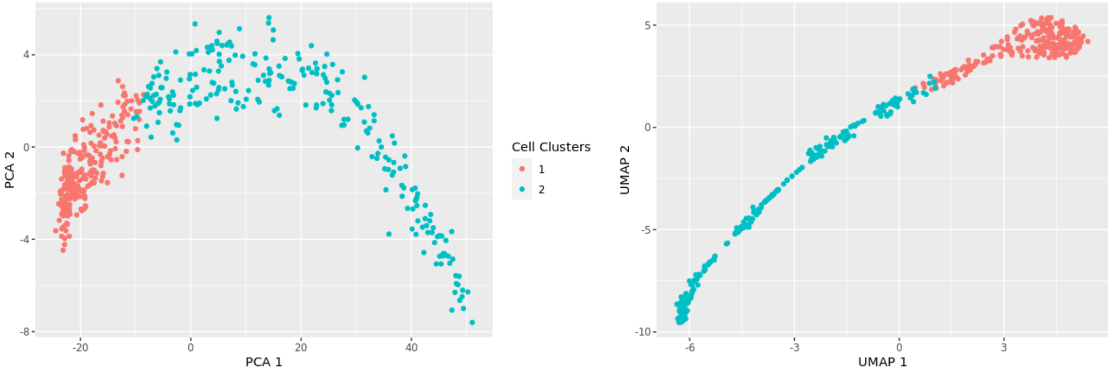

Figure 3.1: PCA and UMAP visualization on original single-lineage data

In our experiment, we found several doublets in the simulated single-lineage data. We detected those doublets by function $doubletThresholding()$ from scDblFinder, which was developed by Germain et al. [2]. About 89 out of 500 cells are recognized as doublets from the function result. Here are the PCA and UMAP plots of the original simulated data with doublet scores being highlighted:

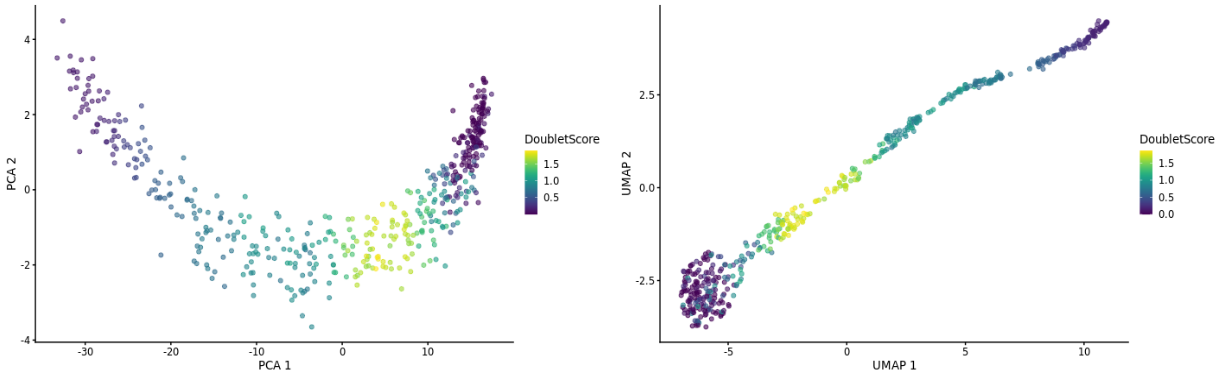

Figure 3.2: Doublet detection on original single-lineage data

where the doublet score for each cell is a measure of the doublet density near that cell.

Then we eliminated about 57% of the detected doublets and produced the doublet reduced dataset for the experiment. Only 38 out of 432 cells are recognized as doublets in the reduced dataset. Here are the PCA and UMAP plots of the doublet reduced dataset with doublet scores being highlighted:

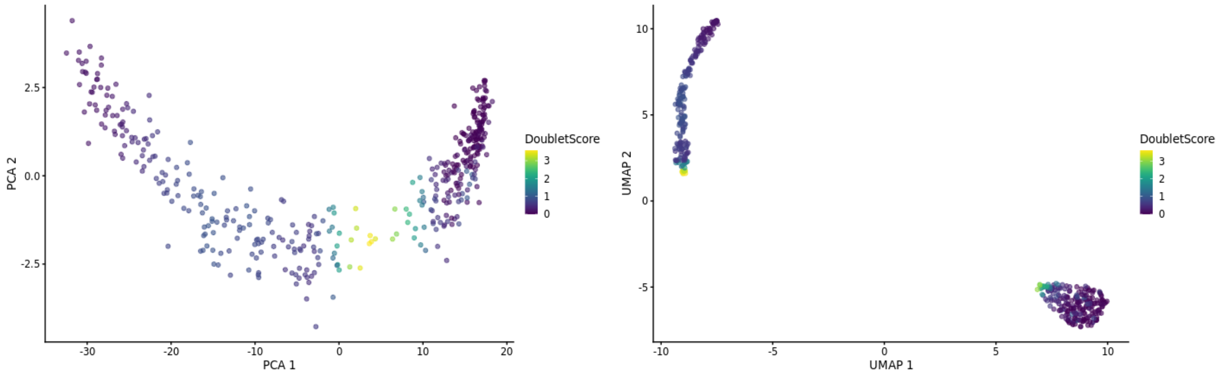

Figure 3.3: Doublet detection on doublet reduced single-lineage data

We also obtaind the PCA and UMAP plots of the doublet reduced dataset with two cell clusters:

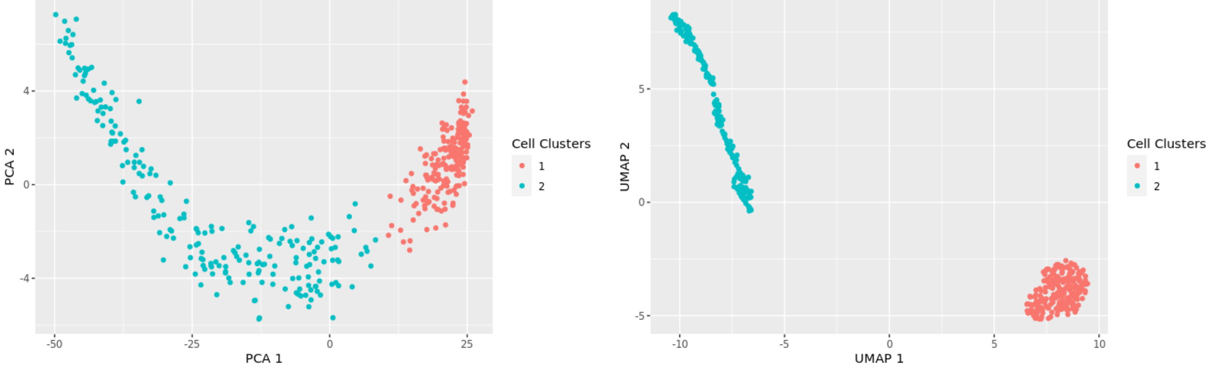

Figure 3.4: PCA and UMAP visualization on doublet reduced single-lineage data

3.1.2 Bifurcation Dataset

For the bifurcation data, we borrowed an existing single-cell data simulator named dyntoy [10, 14] to generate the datasets. After the simulation, we obtained 1000 genes and 1000 cells; half of the genes are DE genes. The cells in the simulated dataset are separated into four clusters: M1, M2, M3, and M4. We also used the slingshot to make the trajectory inference on the bifurcation data. Since it is a bifurcation dataset, two lineages have been inferred from slingshot: $M1 \to M3 \to M2$ and $M1 \to M3 \to M4$. Here are the PCA and UMAP plots of the original bifurcation data:

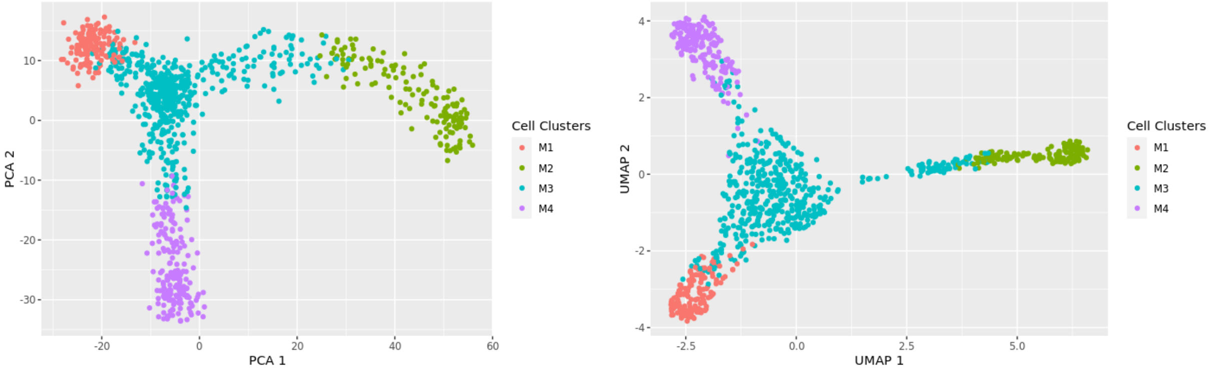

Figure 3.5: PCA and UMAP visualization on original bifurcation data

In the simulated bifurcation data, we also detected doublets by scDblFinder. According to the result, about 196 out of 1000 cells are doublets. Here are the PCA and UMAP plots of original bifurcation data with doublet scores being highlighted:

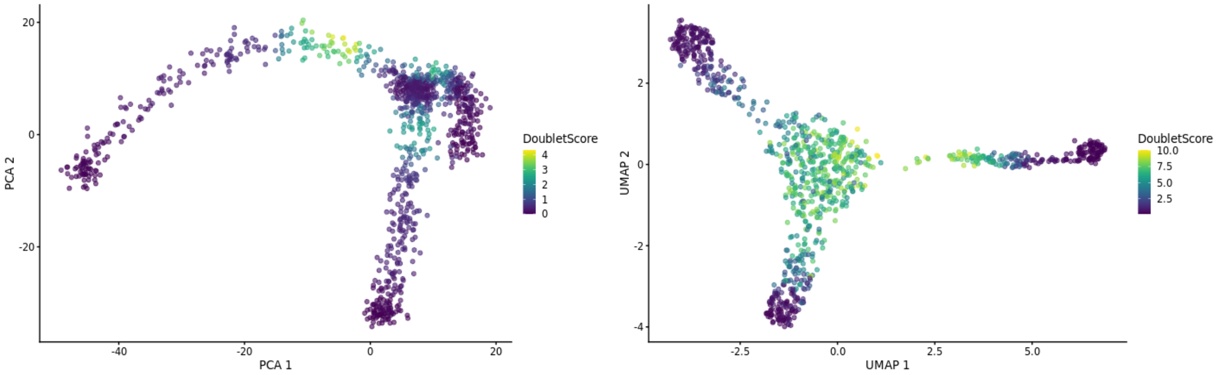

Figure 3.6: Doublet detection on original bifurcation data

Similar to what we did to the single-lineage data, we eliminated about 41% of the doublets in the original bifurcation data. After the reduction, there are about 115 out of 851 cells are doublets in the dataset. Here are the PCA and UMAP plots of the doublet reduced bifurcation data with doublet scores being highlighted:

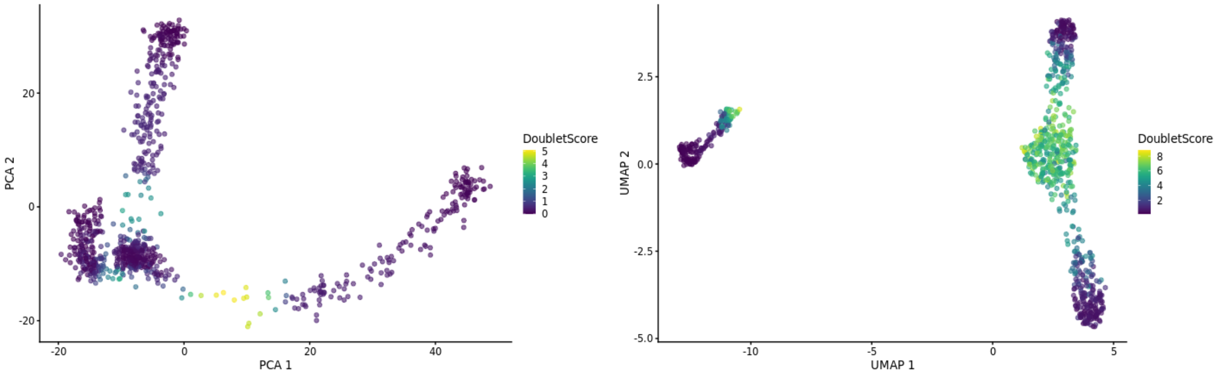

Figure 3.7: Doublet detection on doublet reduced bifurcation data

And we also obtaind the PCA and UMAP plots for doublet reduced bifurcation data with cell clusters:

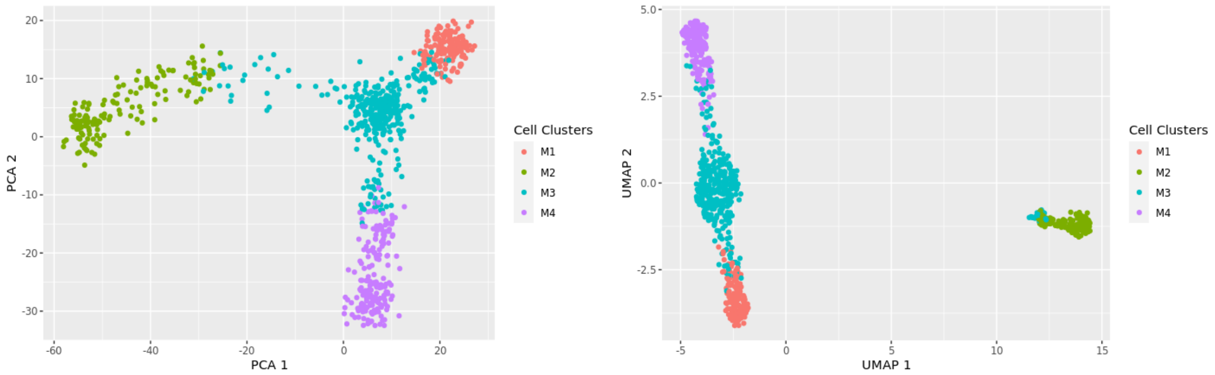

Figure 3.8: PCA and UMAP visualization on doublet reduced bifurcation data

CHAPTER 4

Results

Once we obtaind the simulated datasets, we applied three DE gene inference methods, tradeSeq, PseudotimeDE, and qgamDE, to these datasets. We compared the DE gene detection result among those methods. For both tradeSeq and qgamDE, there is only one set of DE gene detection outputs. For PseudotimeDE, there are two sets of DE gene detection outputs, 'fix' and 'para'; the 'fix' output ignores the uncertainty of inferred pseudotime in the detection process, while the 'para' output considers the uncertainty. The comparative metrics that we included in this thesis are area under the ROC curve (AUC), false discovery rate (FDR), power, and uniformity of the distribution of p-values for non-DE genes.

4.1 Simulated Single-lineage Data

For single-lineage datasets, we obtaind the ROC Curve for DE gene detection from each method (see Fig. 4.1 and Fig. 4.2).

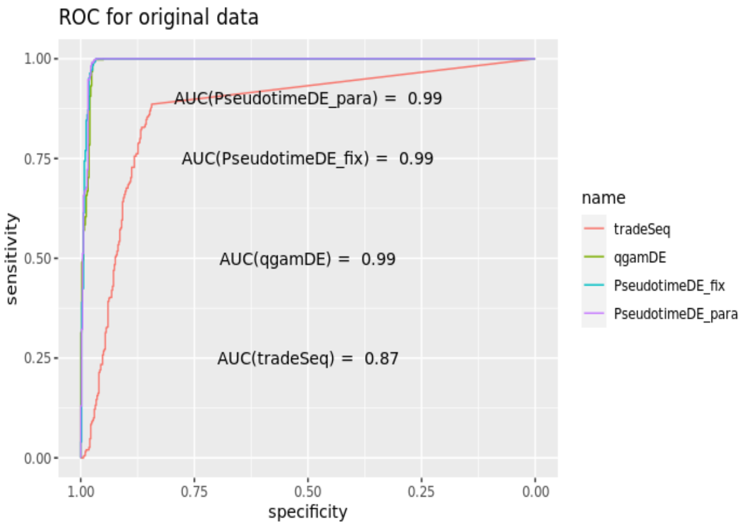

Figure 4.1: ROC of DE gene identification on original single-lineage data

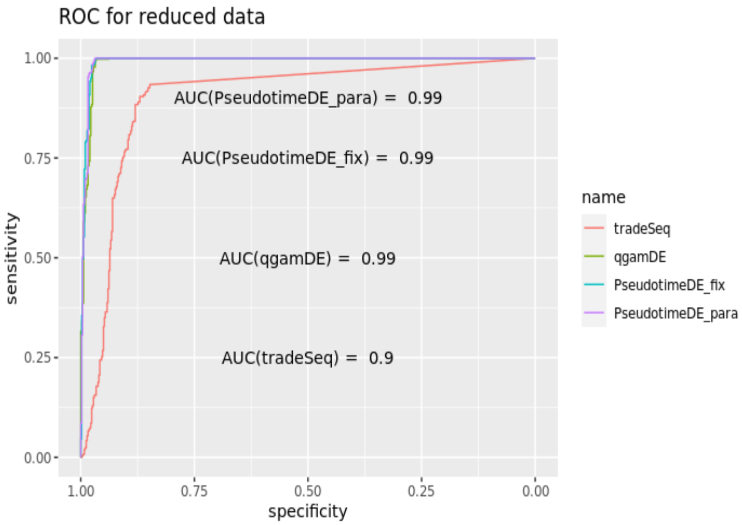

Figure 4.2: ROC of DE gene identification on doublet reduced single-lineage data

From the ROC plot, we can observe that the AUC of tradeSeq increases with doublet reduction. For PseudotimeDE and qgamDE, the AUC does not change by reducing the doublets.

We also compared the FDR and power among the methods (see Fig. 4.3 and Fig. 4.4).

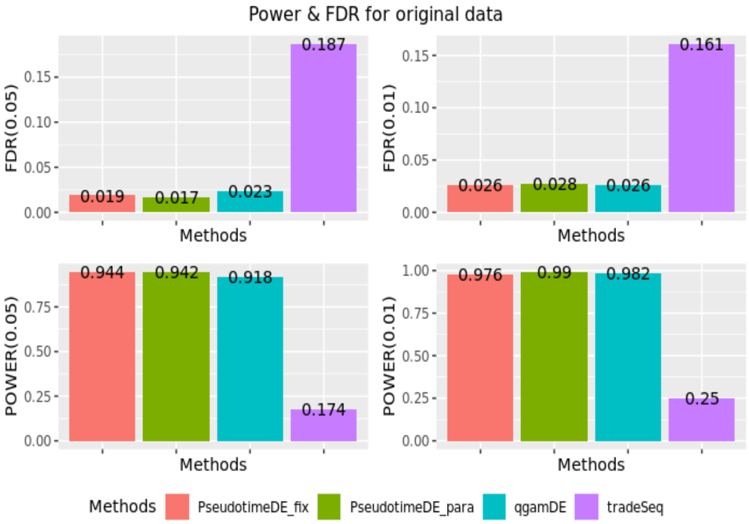

Figure 4.3: FDR and power of DE gene identification on original single-lineage data

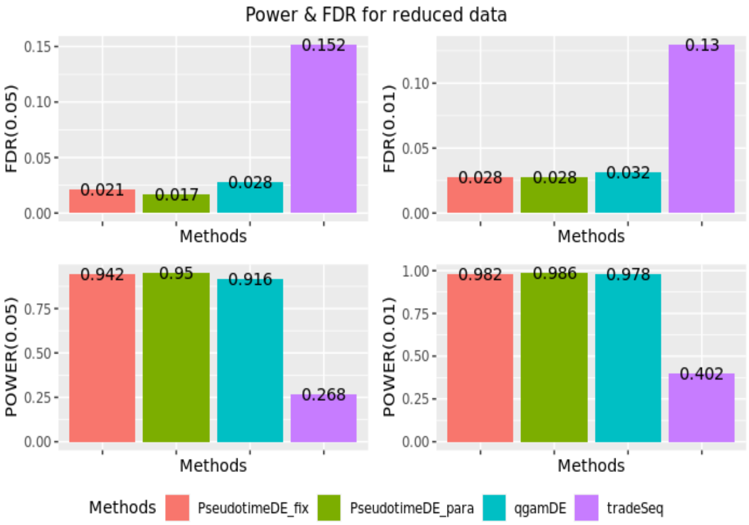

Figure 4.4: FDR and power of DE gene identification on doublet reduced single-lineage data

Also, we still see that the FDR of tradeSeq decreases and the power of tradeSeq increases with doublet reduction, while FDR and power do not change that much for qgamDE and PseudotimeDE.

In the end, we also compared the uniformity of the p-values of non-DE genes. The p-values should be uniformly distributed when the null hypothesis is true (non-DE genes). Here are the p-value distribution plots for both original data and doublet reduced data, with the p-values of the Kolmogorov-Smirnov Test being specified (see Fig. 4.5 and Fig. 4.6).

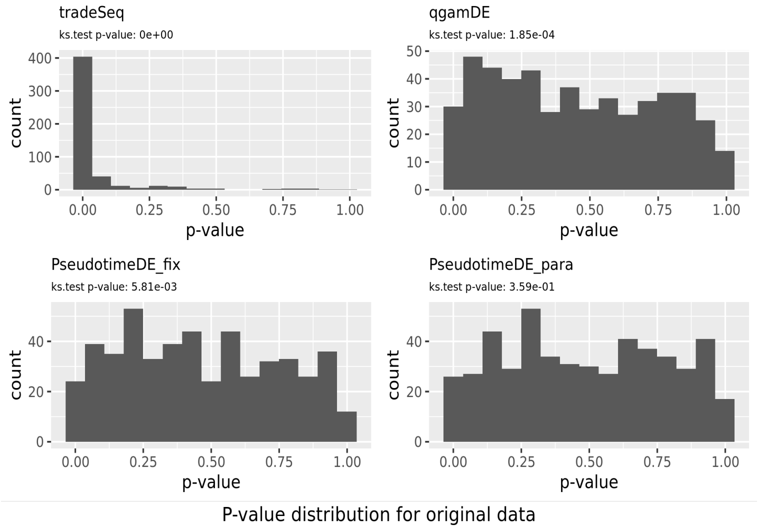

Figure 4.5: P-value distribution on original single-lineage data

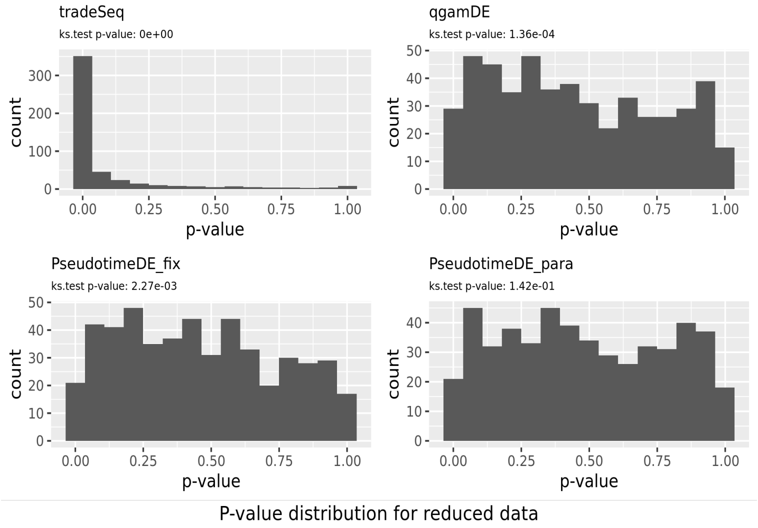

Figure 4.6: P-value distribution on doublet reduced single-lineage data

According to the figures, the p-value distributions from qgamDE and PseudotimeDE are close to the uniform distribution. In contrast, the p-value distributions from tradeSeq are right-skewed for both original and doublet reduced data.

4.2 Simulated Bifurcation Data

There are two lineages within each dataset for the bifurcation data, and then there will be two sets of outputs for each comparative metric. For the ROC curve, we obtaind ROC curves of DE gene identification for each lineage (see Fig. 4.7 and Fig. 4.8).

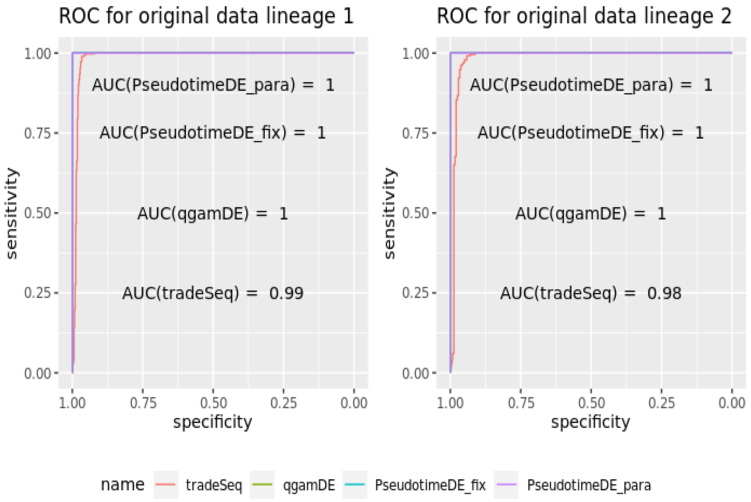

Figure 4.7: ROC of DE gene identification on original bifurcation data

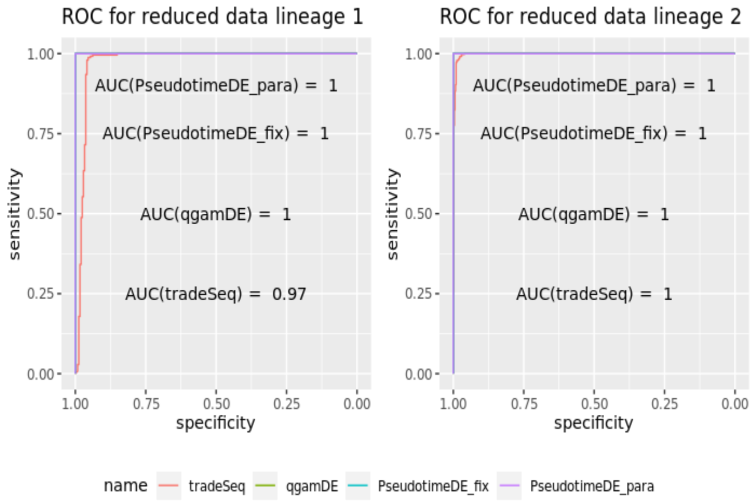

Figure 4.8: ROC of DE gene identification on doublet reduced bifurcation data

The ROC plots show that the performance of tradeSeq is improved after doublet removal. However, the results for qgamDE and PseudotimeDE are accurate and stable. It seems like there are no changes before and after the doublet reduction. And we still obtaind the FDR and power for the three methods (see Fig. 4.9, Fig. 4.10, Fig. 4.11 and Fig. 4.12).

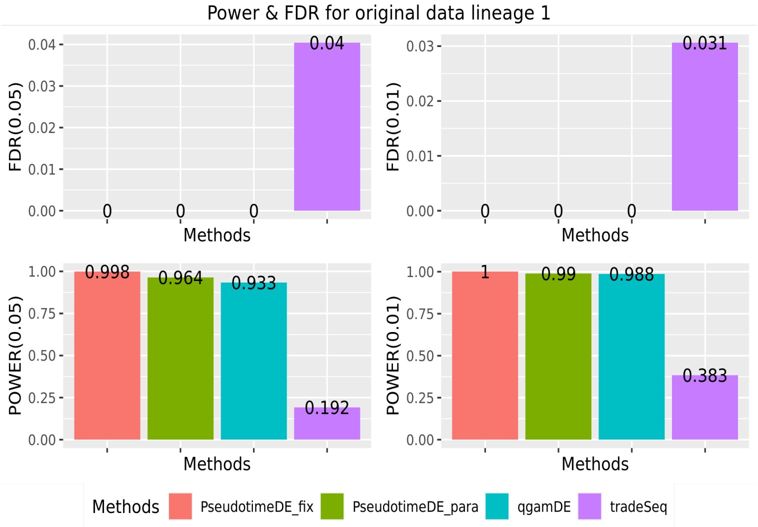

Figure 4.9: FDR and power of DE gene identification of original bifurcation lineage 1

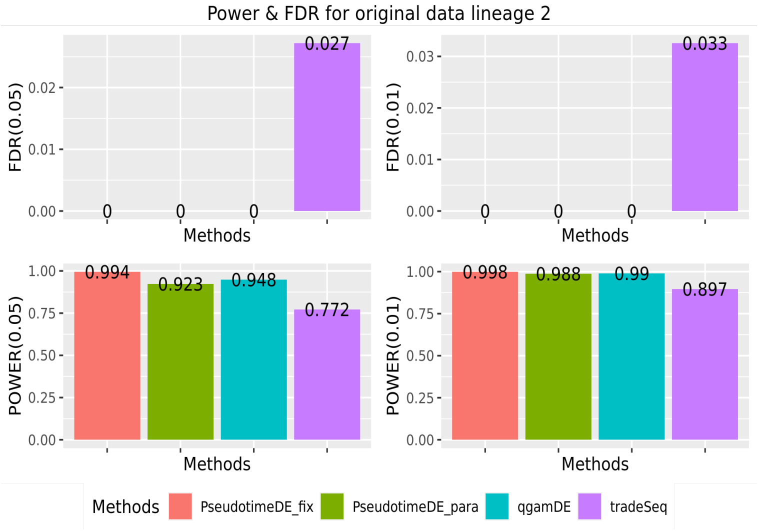

Figure 4.10: FDR and power of DE gene identification of original bifurcation lineage 2

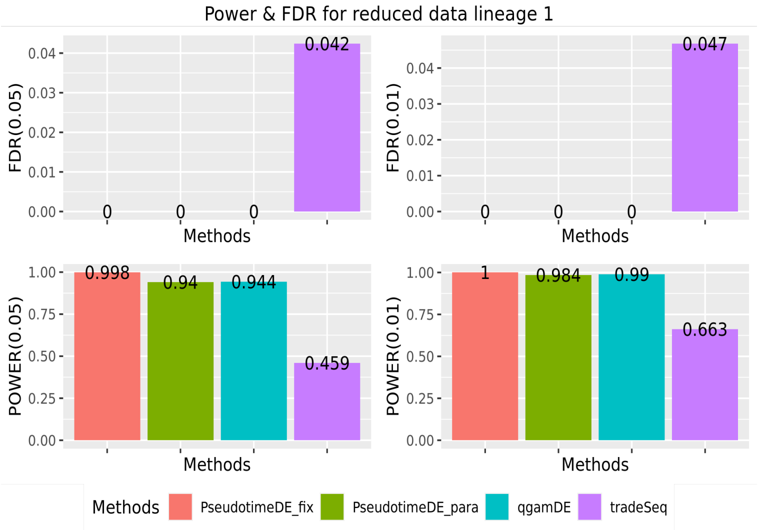

Figure 4.11: FDR and power of DE gene identification of doublet reduced bifurcation lineage 1

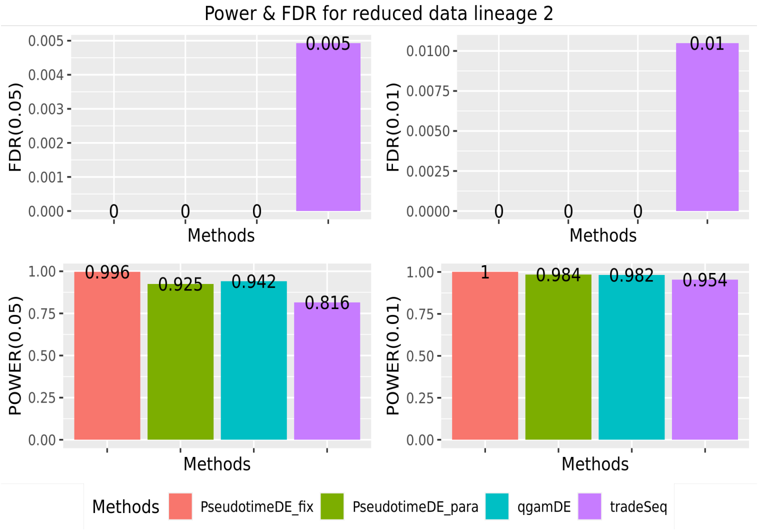

Figure 4.12: FDR and power of DE gene identification of doublet reduced bifurcation lineage 2

The FDR and power plot shows that the FDR decreases for tradeSeq when doublets are removed for the cells in the second lineage. At the same time, the FDR does not change for qgamDE and PseudotimeDE no matter whether the doublets are reduced or not. As for the power, it is obvious that the power for tradeSeq improved when the doublets were removed. However, the changes in power are very tiny for qgamDE and PseudotimeDE.

At last, we also checked the uniformity of the p-values of non-DE genes. Here are the distribution plots of the p-values (see Fig. 4.13, Fig. 4.14, Fig. 4.15 and Fig. 4.16).

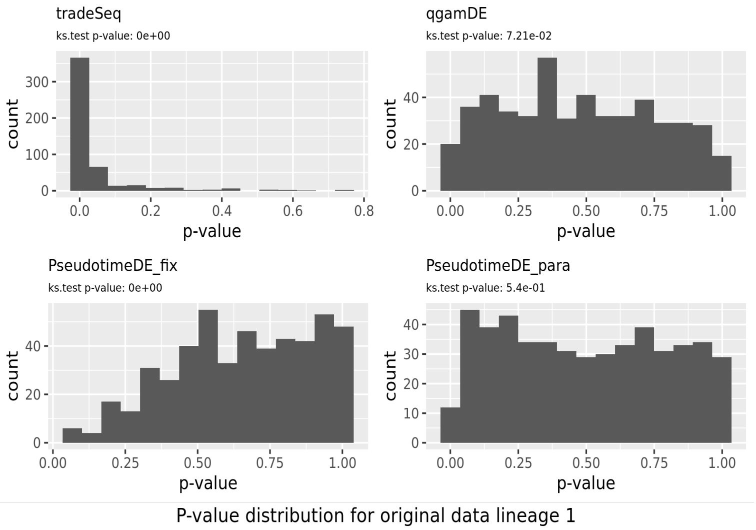

Figure 4.13: P-value distribution of original bifurcation lineage 1

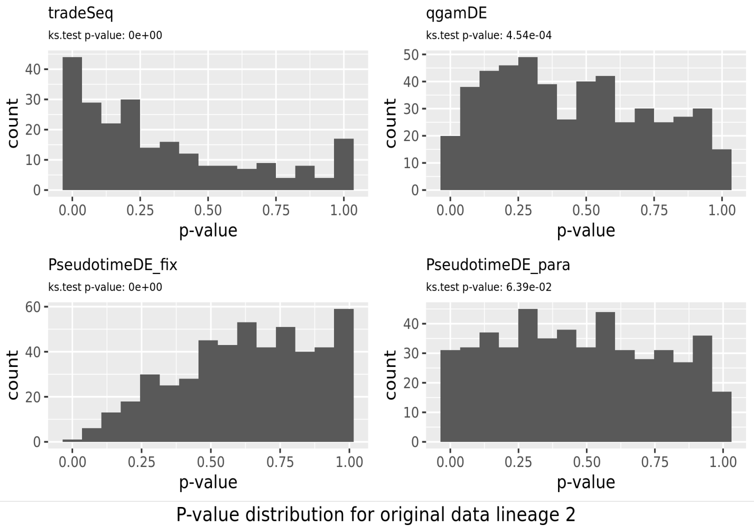

Figure 4.14: P-value distribution of original bifurcation lineage 2

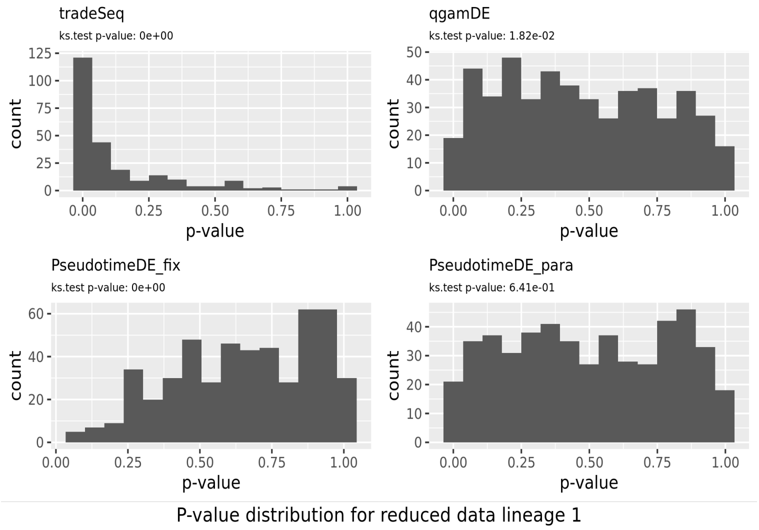

Figure 4.15: P-value distribution of doublet reduced bifurcation lineage 1

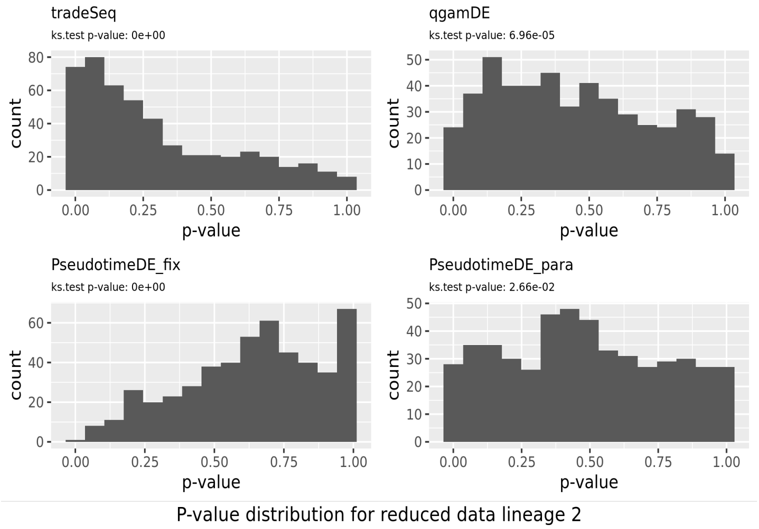

Figure 4.16: P-value distribution of doublet reduced bifurcation lineage 2

According to the figures, the p-value distributions from tradeSeq and PseudotimeDE_fix are either right-skewed or left-skewed. The p-value distributions from qgamDE and PseudotimeDE_para are similar to the uniform distribution.

Overall, PseudotimeDE and qgamDE are better than tradeSeq. One cause of the inferior performance of tradeSeq that we speculate is its problematic p-value calculation. Both PseudotimeDE paper and tradeSeq paper mentioned that the p-values in tradeSeq are calculated based on the chi-square distribution [11, 14]. However, such p-value calculation is not a proper approximation to the null distribution [11]. Based on the uncalibrated p-values, tradeSeq will generate inaccurate results in the p-value-based statistical procedures such as FDR control [11].

CHAPTER 5

Discussion

The key idea behind qgamDE can be summarized as follows: we apply QGAM to fit a model to find the relationship between the gene expression and the pseudotime, and we test the hypothesis of whether the additive terms in the fitted model are 0 to decide if a gene is differentially expressed. If the p-value of this hypothesis is small for a gene, then this gene tends to be DE; otherwise, the gene should be a non-DE gene. In our current experiment, we only fit the 0.5 quantile of the response variable. Moreover, it is worth exploring the results with different quantiles of interest.

Although the performance of qgamDE seems good in our experiment, its p-value of gene j is theoretically problematic. qgamDE directly puts $\hat{f}_j$ from QGAM into S_j, and the p-value is generated without calibration. To figure out why qgamDE performs well even if its p-values are uncalibrated needs further exploration. We speculate that such a situation is related to the fact that QGAM fits the median of the response variable during the DE gene inference process.

The method of qgamDE has been added to the PseudotimeDE package as new functionality. In PseudotimeDE (QGAM), users can set the quantile they want to fit to infer the relationship between gene expression and pseudotime. In the PseudotimeDE (QGAM) vignettes, we apply the inference method to a Smart-seq dataset containing primary mouse dendritic cells (DCs) stimulated with lipopolysaccharide (LPS). The original dataset can be found at Gene Expression Omnibus (GEO) under accession ID GSE45719 [11]. Here are the plots of

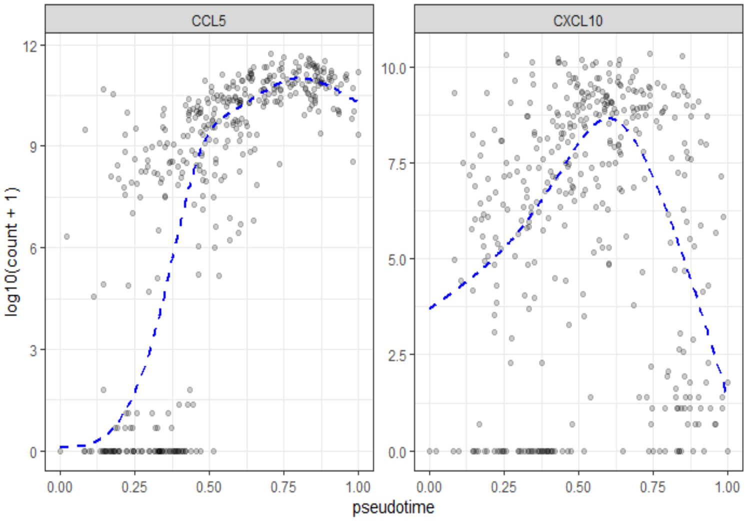

Figure 5.2: QGAM fitting with 0.5 quantile of interest

estimated gene trajectories for gene CCL5 and gene CXCL10 with 0.1, 0.5, and 0.9 quantiles:

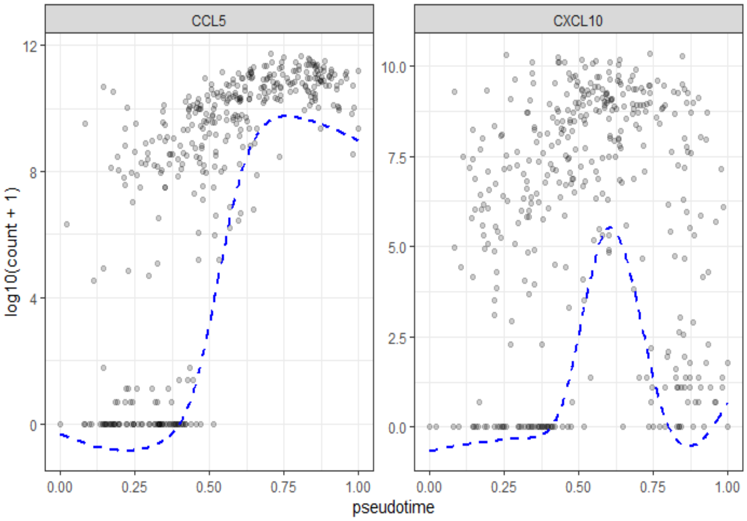

Figure 5.1: QGAM fitting with 0.1 quantile of interest

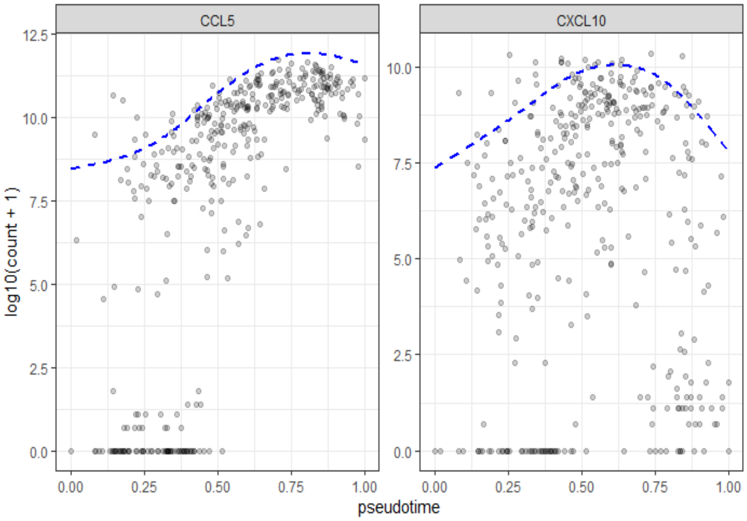

Figure 5.3: QGAM fitting with 0.9 quantile of interest

Our tool can correctly recognize these two genes as DE genes. Besides trying other quantiles of interest, we can also consider the zero-inflation problem. This problem is about the controversy of whether excess zeros that Poisson or negative binomial distributions cannot explain are biological meaningful or not [11]. In the GAM-based PseudotimeDE, this issue was overcome by introducing both negative binomial GAM and zero-inflated negative binomial GAM. The former treats excess zeros as biologically meaningful, and the latter does not [11]. In such a way, users can choose the model to do the DE gene inference task accordingly. It is worth seeing if it is possible to let PseudotimeDE QGAM take care of the zero-inflation issue.

CHAPTER 6

Conclusion

This thesis introduced a new application of QGAM, DE gene inference. The main difference between qgamDE and the state-of-the-art DE gene inference methods is that qgamDE used the quantile generalized additive model instead of the generalized additive model to make the inference. After we applied both qgamDE and the other two methods to the simulated scRNA-seq datasets, the experiment showed that qgamDE could produce a relatively accurate and stable result regardless of the number of outliers. However, the current results are limited because they are all simulation-based. Real scRNA-seq data must be used to check the practical efficacy of qgamDE.

REFERENCES

[1] Matteo Fasiolo, Simon N Wood, Margaux Zaffran, Raphaël Nedellec, and Yannig Goude. Fast calibrated additive quantile regression. *Journal of the American Statistical Association*, 116(535):1402–1412, 2021.

[2] Pierre-Luc Germain, Aaron Lun, Will Macnair, and Mark D Robinson. Doublet identification in single-cell sequencing data using scdblfinder. *F1000Research*, 10(979):979, 2021.

[3] Ashraful Haque, Jessica A. Engel, Sarah A. Teichmann, and Tapio Lönnberg. A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome Medicine*, 9, 2017.

[4] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

[5] Zhicheng Ji and Hongkai Ji. Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117, 2016.

[6] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell rna-seq data. *Nature Reviews Genetics*, 20(5):273–282, 2019.

[7] Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50, 1978.

[8] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.

[9] Belinda Phipson and Gordon K Smyth. Permutation p-values should never be zero: calculating exact p-values when permutations are randomly drawn. *Statistical applications in genetics and molecular biology*, 9(1), 2010.

[10] Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

[11] Dongyuan Song and Jingyi Jessica Li. Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data. *Genome biology*, 22(1):1–25, 2021.

[12] Kelly Street, Davide Risso, Russell B Fletcher, Diya Das, John Ngai, Nir Yosef, Elizabeth Purdom, and Sandrine Dudoit. Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16, 2018.

[13] Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

[14] Koen Van den Berge, Hector Roux de Bézieux, Kelly Street, Wouter Saelens, Robrecht Cannoodt, Yvan Saeys, Sandrine Dudoit, and Lieven Clement. Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications*, 11(1):1–13, 2020.

[15] Samuel L Wolock, Romain Lopez, and Allon M Klein. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *Cell systems*, 8(4):281–291, 2019.

[16] Simon N Wood. On p-values for smooth components of an extended generalized additive model. *Biometrika*, 100(1):221–228, 2013.

[17] Grace XY Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, et al. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8(1):1–12, 2017.