Applications of generalized additive models and copulas to single-cell RNA-seq computational method development

Jingyi Jessica Li
Department of Statistics
University of California, Los Angeles
http://jsb.ucla.edu
1. **PseudotimeDE**: inference of differential gene expression along cell pseudotime with valid p-values from single-cell RNA-seq data
   - by **Dongyuan Song** (宋东源; 2nd-year Bioinformatics PhD student)
   - *Genome Biology*

   ![PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data](https://doi.org/10.1101/2020.11.17.387795)

2. **scDesign2**: a transparent simulator that generates realistic single-cell gene expression count data with gene correlations captured
   - by **Tianyi Sun** (孙天毅; 4th-year Statistics PhD student) *et al.*
   - accepted by *RECOMB* and in press at *Genome Biology*
   - bioRxiv: https://doi.org/10.1101/2020.11.17.387795
PseudotimeDE
Pseudotime inference

- **Pseudotime**: a latent “temporal” variable that reflects a cell’s relative transcriptome status among all cells
- **Pseudotime inference** (trajectory inference): estimate the pseudotime of cells, i.e., order cells along a trajectory (lineage) based on transcriptome similarities
- Popular methods:
  - Monocle3 *(Trapnell et al. 2014)*
  - TSCAN *(Ji et al. 2016)*
  - Slingshot *(Street et al. 2018)*
Differential gene expression along cell pseudotime

- Differentially expressed (DE) gene: a gene whose expected expression changes along cell pseudotime
- Question: how to identify DE genes?
Limitations of existing methods

- **tradeSeq** (Van den Berge *et al.* 2020)
  - Test if a gene is DE based on a generalized additive model (GAM) (Hastie and Tibshirani, 1986, 1990)

- **Monocle3** (Trapnell *et al.* 2014)
  - Test if a gene is DE based on a generalized linear model (GLM) (McCullagh, 1983)

- Both methods are regression-based:
  - response: a gene’s expression level in a cell
  - predictor/covariate: a cell’s pseudotime

- **Limitation**: cell pseudotime is treated as fixed with uncertainty ignored

- **Why is cell pseudotime random?**
  - pseudotime is not observed but inferred; inference involves uncertainty

- **This ignorance of pseudotime uncertainty may result in invalid \( p \)-values**
Pseudotime inference uncertainty

(a) PC1 vs PC2 for different pseudotimes of subsamples. (b) UMAP 1 vs UMAP 2 for pseudotimes of subsamples.
Our proposal: PseudotimeDE

Subsampling cells

Pseudotime inference

Permutation on cells

NB/ZINB-GAM fitted to gene $j$

Null distribution of $S_j$

Test statistic $S_j$

$p$-value of gene $j$
PseudotimeDE: notations

- $\mathbf{Y} = (Y_{ij})$: an $n \times m$ gene expression count matrix ($n$ cells and $m$ genes)

- $\mathbf{T} = (T_1, \ldots, T_i, \ldots, T_n)^T$: cell pseudotime inferred from $\mathbf{Y}$

- To capture the uncertainty of pseudotime $\mathbf{T}$, we subsample 80% cells in $\mathbf{Y}$ for $B$ times (default $B = 1000$); in the $b$-th subsample:
  - $\mathbf{Y}^b = (Y_{ij}^b)$, an $n' \times m$ matrix where $n' = \lfloor 0.8n \rfloor$
  - $\mathbf{T}^b = (T_1^b, \ldots, T_{n'}^b)^T$: cell pseudotime inferred from $\mathbf{Y}^b$
  - $\mathbf{T}^{*b} = (T_1^{*b}, \ldots, T_{n'}^{*b})^T$: permuted cell pseudotime
PseudotimeDE: GAM

- Negative-Binomial Generalized Additive Model (NB-GAM)

\[
\begin{align*}
Y_{ij} &\sim NB(\mu_{ij}, \phi_j) \\
\log(\mu_{ij}) &= \beta_{j0} + f_j(T_i)
\end{align*}
\]

- Zero-Inflated Negative-Binomial Generalized Additive Model (ZINB-GAM)

\[
\begin{align*}
Z_{ij} &\sim Ber(p_{ij}) \\
Y_{ij} | Z_{ij} &\sim Z_{ij} \cdot NB(\mu_{ij}, \phi_j) + (1 - Z_{ij}) \cdot 0 \\
\log(\mu_{ij}) &= \beta_{j0} + f_j(T_i) \\
\logit(p_{ij}) &= \alpha_{j0} + \alpha_{j1} \log(\mu_{ij})
\end{align*}
\]

where \( f_j(T_i) = \sum_{k=1}^{K} b_k(T_i) \beta_{jk} \) is a cubic spline function
NB-GAM (blue) v.s. ZINB-GAM (red)
PseudotimeDE: statistical test

- Null and alternative hypotheses for gene $j$:

  $H_0 : f_j(\cdot) = 0$ vs. $H_1 : f_j(\cdot) \neq 0$

- Fit GAM to $Y$ and $T$. Denote the estimate of $(f_j(T_1), \ldots, f_j(T_n))^T$ by $\hat{f}_j$ and estimated covariance matrix of $\hat{f}_j$ by $\hat{V}_{f_j}$

- Test statistic:

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

  where $\hat{V}_{f_j}^{-r}$ is the rank-$r$ pseudo-inverse of $\hat{V}_{f_j}$

- Observed value of $S_j$ denoted by $s_j$

- For $b = 1, \ldots, B$, fit GAM to $Y^b$ and $T^{*b}$; calculate the test statistic $s^b_j$

- $\{s^1_j, \ldots, s^B_j\}$: null values of the test statistic $S_j$

- **Gene $j$’s p-value** $p_j \leftarrow s_j, \{s^1_j, \ldots, s^B_j\}$
Real data example: dendritic cells stimulated with LPS

(a) Slingshot

(b) Venn diagram for GO terms:
- PseudotimeDE
- tradeseq
- Monocle3-DE

(c) Bar chart showing the number of significant GO terms:
- PseudotimeDE vs tradeseq
- PseudotimeDE vs Monocle3-DE

(d) Monocle3-PI

(e) Venn diagram for GO terms:
- PseudotimeDE
- tradeseq
- Monocle3-DE

(f) Bar chart showing the number of significant GO terms:
- PseudotimeDE vs tradeseq
- PseudotimeDE vs Monocle3-DE

(g) Bar chart showing the number of significant GO terms:
- PseudotimeDE vs tradeseq
- PseudotimeDE vs Monocle3-DE

Each chart illustrates the comparison of PseudotimeDE, tradeseq, and Monocle3-DE in the context of observed p-values and the count of significant GO terms.
Real data example: dendritic cells stimulated with LPS

PseudotimeDE vs tradeSeq

<table>
<thead>
<tr>
<th>ID</th>
<th>description</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0043410</td>
<td>positive regulation of MAPK cascade</td>
<td>0.00331</td>
</tr>
<tr>
<td>GO:0050729</td>
<td>positive regulation of inflammatory response</td>
<td>0.00424</td>
</tr>
<tr>
<td>GO:0071222</td>
<td>cellular response to lipopolysaccharide</td>
<td>0.00609</td>
</tr>
</tbody>
</table>

PseudotimeDE vs Monocle3–DE

<table>
<thead>
<tr>
<th>ID</th>
<th>description</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006954</td>
<td>inflammatory response</td>
<td>0.00032</td>
</tr>
<tr>
<td>GO:0043410</td>
<td>positive regulation of MAPK cascade</td>
<td>0.00079</td>
</tr>
<tr>
<td>GO:0042742</td>
<td>defense response to bacterium</td>
<td>0.00573</td>
</tr>
</tbody>
</table>

PseudotimeDE vs tradeSeq

<table>
<thead>
<tr>
<th>ID</th>
<th>description</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0002250</td>
<td>adaptive immune response</td>
<td>0.00812</td>
</tr>
<tr>
<td>GO:0002673</td>
<td>regulation of acute inflammatory response</td>
<td>0.00955</td>
</tr>
<tr>
<td>GO:0017001</td>
<td>antibiotic catabolic process</td>
<td>0.00955</td>
</tr>
</tbody>
</table>

PseudotimeDE vs Monocle3–DE

<table>
<thead>
<tr>
<th>ID</th>
<th>description</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0045087</td>
<td>innate immune response</td>
<td>0.00016</td>
</tr>
<tr>
<td>GO:0009617</td>
<td>response to bacterium</td>
<td>0.00039</td>
</tr>
<tr>
<td>GO:0050829</td>
<td>defense response to Gram–negative bacterium</td>
<td>0.00418</td>
</tr>
</tbody>
</table>
Learning more about GAMs

- Book: *Generalized Additive Models: an introduction with R* by Dr. Simon Woods

- R package mgcv by Dr. Simon Woods
scDesign2
Motivation

- **Experimental design:**
  - How to choose among existing experimental protocols?
  - Given a chosen protocol, how to determine the optimal parameters for the experiment (cell number and seq. depth)?

- **Computational benchmarking:**
  - How to choose among available computational methods for data analysis?

- **Use a realistic simulator to answer these questions!**
## Existing scRNA-seq simulators

<table>
<thead>
<tr>
<th>Simulator</th>
<th>Property</th>
<th>protocol adaptive</th>
<th>genes preserved</th>
<th>gene cor. captured</th>
<th>cell num. seq. depth flexible</th>
<th>easy to interpret</th>
<th>comp. &amp; sample efficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>dyngen</td>
<td></td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Lun2</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>powsimR</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>PROSST</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>scDD</td>
<td></td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>scDesign</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>scGAN</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>splat simple</td>
<td></td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>splat</td>
<td></td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>kersplat</td>
<td></td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SPARSim</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SymSim</td>
<td></td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>ZINB-WaVE</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>SPsimSeq</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Our proposal: scDesign2

Key Features of scDesign2:

• scDesign2 can reliably capture gene correlations
• scDesign2 can simulate data with varying cell numbers and sequencing depths, facilitating the design of experiments

Input

Parameter Estimation

Data Simulation

Real count matrix

Sub-matrices

Gene joint distribution

User specified sequencing depth & number of cells

(e.g. sequencing depth becomes lower)

Guidance for Experimental Design

Evaluation of Computational Methods
scDesign2: notations

- Denote the scRNA-seq count matrix as $\mathbf{X} \in \mathbb{N}^{p \times n}$, with $p$ genes and $n$ cells.

- Assume that $\mathbf{X}$ contains $K$ cell types and the cell memberships are known in advance.

- Suppose there are $n^{(k)}$ cells in cell type $k$, $k = 1, \ldots, K$, and denote the count matrix for cell type $k$ as $\mathbf{X}^{(k)}$.

- Our goal is to fit a parametric, probabilistic model of all genes' expression in each cell type $k$.

- For simplicity of notation, we drop the subscript $k$ in the following discussion.
scDesign2: marginal distribution of each gene $i$

- Model counts directly

- Denote $X_j = (X_{1j}, \ldots, X_{pj}) \in \mathbb{N}^p$ as the gene expression vector for cell $j$, $j = 1, \ldots, n$. We assume that the $X_j$’s are i.i.d. — $p$ variables; $n$ observations

- $x_{ij}$: observed count of gene $i$ in cell $j$

- Select a marginal count distribution for gene $i$’s count $X_{ij}$ from Poisson, zero-inflated Poisson, negative binomial, and zero-inflated negative binomial
scDesign2: joint distribution of all genes

- Use the copula framework

- Denote $F : \mathbb{N}^p \rightarrow [0, 1]$ as the joint cumulative distribution function (CDF) of $X_j \in \mathbb{N}^p$ and $F_i : \mathbb{N} \rightarrow [0, 1]$ as the marginal CDF of $X_{ij}$

- By Sklar’s theorem [Sklar 1959], there exists a copula function $C : [0, 1]^p \rightarrow [0, 1]$ such that

\[ F(x_{1j}, \ldots, x_{pj}) = C(F_1(x_{1j}), \ldots, F_p(x_{pj})) \]

- The copula function $C(\cdot)$ is unique for continuous distributions, but not for discrete distributions (unidentifiable) [Genest et al 2007]
scDesign2: distributional transform and the Gaussian copula

- **Distributional transform**: necessary for discrete variable [Rüschendorf 2013].
  - Sample $v_{ij}$ from $\text{Uniform}[0, 1]$ independently for $i = 1, \ldots, p$ and $j = 1, \ldots, n$
  - Calculate $u_{ij}$ as
    \[
    u_{ij} = v_{ij} F_i(x_{ij} - 1) + (1 - v_{ij}) F_i(x_{ij})
    \]

- **Gaussian copula**: Denote $\Phi$ as the CDF of a standard Gaussian random variable, we can express the joint distribution of $X_j$ as
  \[
  F(x_{1j}, \ldots, x_{pj}) = \Phi_p(\Phi^{-1}(u_{1j}), \ldots, \Phi^{-1}(u_{pj})| R)
  \]
  where $\Phi_p(\cdot|R)$ is a joint Gaussian CDF with a zero mean vector and a covariance matrix that is equal to the correlation matrix $R$
scDesign2: joint distribution fitting

- Denote $\hat{F}_i$ as the estimated marginal distribution of gene $i$

- Sample $v_{ij}$ from Uniform$[0, 1]$ independently for $i = 1, \ldots, p$ and $j = 1, \ldots, n$

- Calculate $u_{ij}$ as

$$u_{ij} = v_{ij}\hat{F}_i(x_{ij} - 1) + (1 - v_{ij})\hat{F}_i(x_{ij})$$

- Calculate $\hat{R}$ as the sample correlation matrix of $(\Phi^{-1}(u_{1j}), \ldots, \Phi^{-1}(u_{pj}))^T$, $j = 1, \ldots, n$
scDesign2: data simulation

- **Input from previous step:**
  - fitted joint gene distributions (one per cell type)
  - cell type proportions

- **User-specified input:**
  - number of cells to simulate
  - total sequencing depth

- **Output:**
  - a synthetic gene-by-cell count matrix with $K$ cell types
  - fitted model parameters (optional)
scDesign2 vs. existing scRNA-seq simulators

Data: mouse small intestinal goblet cells by 10x Genomics [Haber et al., Nature (2017)]
scDesign2 vs. existing scRNA-seq simulators

Data: dendrocytes subtype 1 of human blood by Smart-Seq2 [Villani et al., Science (2017)]
scDesign2 vs. existing scRNA-seq simulators

Data: mouse small intestinal epithelium cells by 10x Genomics [Haber et al., Nature (2017)]
Application 1: simulation for other single-cell technologies

Data: mouse hypothalamic preoptic region by MERFISH [Moffitt et al., Science (2018)]
Application 2: benchmarking cell clustering methods

Data: mouse small intestinal epithelium cells by 10x Genomics [Haber et al., Nature (2017)]
Application 3: benchmarking rare cell type detection methods

Data: mouse small intestinal epithelium cells by 10x Genomics [Haber et al., Nature (2017)]
Learning more about copulas

- Book: *Introduction to copulas* by Dr. Roger B Nelson
Summary

- **PseudotimeDE**: finding DE genes along cell pseudotime
  - Well-calibrated $p$-values (essential for FDR control and GSEA)
  - Powerful (thanks to GAM)
  - R package: [https://github.com/SONGDONGYUAN1994/PseudotimeDE](https://github.com/SONGDONGYUAN1994/PseudotimeDE)

- **scDesign2**: generating realistic synthetic single-cell gene expression data
  - Gene correlations preserved (thanks to copula)
  - Probabilistic, transparent, interpretable
  - R package: [https://github.com/JSB-UCLA/scDesign2](https://github.com/JSB-UCLA/scDesign2)
Acknowledgements

Dongyuan Song 宋东源  
(Ph.D. student, UCLA)

Tianyi Sun 孙天毅  
(Ph.D. student, UCLA)

Dr. Wei Vivian Li 李维  
(former Ph.D. student; assistant professor, Rutgers)