Comparing Models of Subject-Clustered Single-Cell Data

Lee Panter

Audrey Hendricks[1], PhD---------------------(Committee Chair and Advisor)
Stephanie Santorico[1], PhD------------------(Committee Member)
Rhonda Bacher[2], PhD------------------------(Committee Member)

[1]The University of Colorado-Denver
[2]The University of Florida

Introduction

Single-Cell (SC) Basics

“Bulk” Sequencing Methods

- Analyze combined expression from thousands/millions of cells
- Often fail to capture variability within sample
- Measurement accuracy less concerning, and protocol dependencies less influential

SC Sequencing Methods

- Analyze expression measurements specific to individual cells
- Hundreds/thousands of SC measurements used for one “SC sample”

Applications of SC methods

- Detecting values differentially expressed across conditions [1]
- Identifying rare cellular subpopulations [2]

Production of SC data & technology

- Increasingly economical to produce SC data with further sampling integration
- Multiple-source samples enable analysis of source-level relationships
- Developmental phases of integrating multiple subjects/samples into a single data set
- Statistical modeling methods incomplete
- Reliability, accuracy, protocol independence still concerning

Introduction

Motivation

- **What problem am I addressing?**
- **In order to solve the problem, what needs to be done?**
- **What do I do to solve the problem?**

What problem am I addressing?

- Single-cell (SC) data is increasing in prevalence
- SC data with multiple subjects emerging for analysis
- Not clear how to analyze subject level relationships (SLRs)

In order to solve the problem, what needs to be done?

- *Demonstrate*: existing statistical models account for SLRs in SC data
- *Compare*: how each method differs/resembles the others

What do I do to solve the problem?

- Outline five modeling methods and how the models account for SLRs in SC data
- Apply the modeling methods to motivating SC data example
  - *Demonstrate* how (if) the models account for SLRs in the motivating example SC data
  - *Compare* how (if) model frameworks account for SLRs in practice

Model Descriptions

Overview of Selected Models

The Models

1. Linear Model (LM)
2. Linear Model with Fixed Effect for Subject (LM-FE)
3. Linear Mixed Effect Model with Random Intercept for Subject (LMM-RI)
4. Linear Mixed Effect Model with Random Intercept and Random Slope for Subject (LMM-RS)
5. Generalized Estimating Equations (GEE)

Model Descriptions

Notation

$$\left(\boldsymbol{X_{ij}}, \boldsymbol{Y_{ij}}\right)$$

subject level predictor-response pair

$$\boldsymbol{i = 1, \ldots N}$$

subject from which measurement was taken

$$\boldsymbol{j = 1, \ldots, n_i}$$

measurement index taken within subject **i**

(repeated measure index)

Overview of Selected Models

Linear Model (LM)

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}}$$

Terms:

- β_0: Intercept
- β_1: Fixed effect slope
- ϵ_{ij}: Residual Error $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

Linear Model (LM) Further Information

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 X_{ij} + \epsilon_{ij}}$$

- Does not account for subject level associations in the data
- Assumes observations are independent
- $\boldsymbol{\beta_1}$ parameter interpreted as population average representation of relationship between predictor and response.
- Nested within all other models

Overview of Selected Models

Linear Model with Fixed Effect (LM-FE)

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_{1i}(subject_j) + \beta_2 X_{ij} + \epsilon_{ij}}$$

Terms:

$$\beta_{1i}(subject_j) = \begin{cases} \beta_{1i} & if\ i = j \\ 0 & if\ i \neq j \end{cases} \quad for\ i = 2, \ldots, N$$

- β_0: Intercept
- β_1: Fixed Effect Intercept (subject)
- β_2: Fixed Effect Slope
- ϵ_{ij}: Residual Error $\quad \epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

Linear Model with Fixed Effect (LM-FE) Further Information

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_{1i}(subject_j) + \beta_2 X_{ij} + \epsilon_{ij}}$$

- Accounts for subject-level associations by:
  - Uniformly shifting the mean of the fitted values specific to a subject
  - Adds N-1 parameters
- Assumes that observations are independent
- $\boldsymbol{\beta_1}$ parameter interpreted as population average representation of relationship between predictor and response, having accounted for average deviation of each subject

Overview of Selected Models

Linear Mixed Model with Random Effect Intercept (LMM-RI)

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}(subject_j) + \epsilon_{ij}}$$

Terms:

- β_0: Intercept
- β_2: Fixed Effect Slope
- b_{0i}: Random Effect Intercept (subject) $b_{0i} \sim N(0, \sigma_b^2)$
- ϵ_{ij}: Residual Error $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

Overview of Selected Models

Linear Mixed Model with Random Effect Intercept and Slope (LMM-RS)

Overview of Selected Models

Generalized Estimating Equations (GEE)

Motivating Example

Data

Initial Data:

- Population: 45 Lupus Nephritis Cases vs 25 Control
- Population: 27 subjects, case/control status not present.
- 9560 SC observations
- Over $3.8 * 10^8$ RNA sequencing (scRNA-seq) variable measures
- 23 Flow Cytometry variables
- 10 metadata variables (subject, cell-type)

Quality Control Data

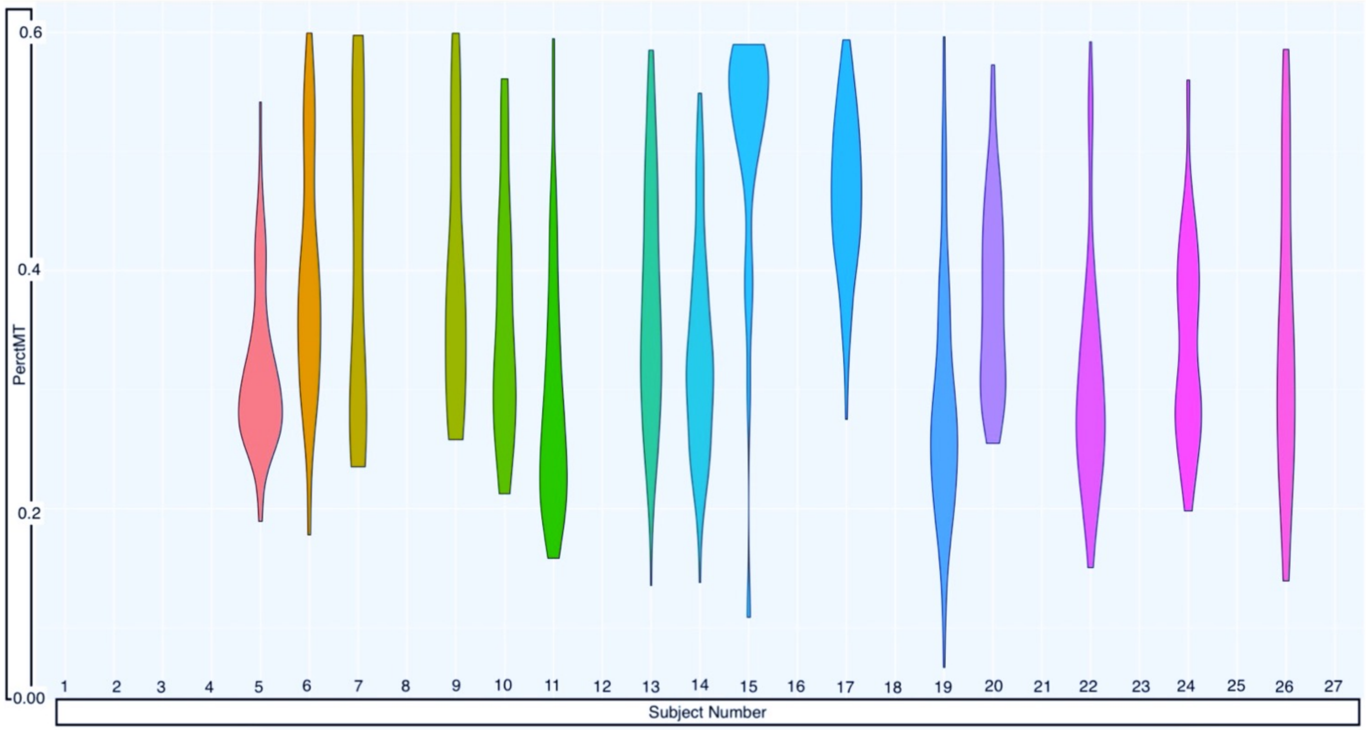
- 15 Subjects
- 1110 SC Observations
- 2 log-transformed scRNA-seq Predictor-Response Pairs

Data Source: 2018 article: "The immune cell landscape in kidneys with Lupus Nephritis patients" [3]

Motivating Example

Models

Proposal: A *method* for estimating subject level associations in SC *data*

- Data Requirements:
  - Single-cell level variable measurements
    - Data is scRNA-seq expression
  - Detectable, subject-level associations between predictor and outcome
    - 

Results

Model Parameters

Model Designation	Model Description	Variable Pair #1: MALAT1<--CD19				Variable Pair #1: FBLN1<--CD34			
		Estimate	Std. Error	Test Statistic	p-value	Estimate	Std. Error	Test Statistic	p-value
LM	Linear Model	4.918e-2	1.455e-2	3.381	7.47e-4	7.884e-1	4.92e-2	4.002	<2e-16
LM-FE	Linear Model with Fixed-Effect Intercept	4.833e-2	1.381e-2	3.500	4.84e-4	1.31e-1	3.42e-2	3.818	1.42e-4
LMM-RI	Linear Mixed Model with Random Intercept	4.920e-2	1.374e-2	3.579	3.6e-4	1.35e-1	3.42e-2	3.95	8.4e-5
LMM-RS	Linear Mixed Model with Random Slope	5.938e-2	3.538e-2	1.678	1.19e-1	1.705e-1	7.29e-2	2.34	6.7e-2
GEE	Generalized Estimating Equations	4.918e-2	1.455e-2	3.381**	7.47e-4	7.884e-1	4.92e-2	4.002**	< 2e-16

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-1.7283	0.0407	20.7401	0.0000
LM-FE	1.7587	0	1.8001	22.8636	1.7587
LMM-RI	-0.0407	-1.7683	0	20.6911	-0.0407
LMM-RS	-17.1775	-18.6090	-17.1438	0	-17.1775
GEE	0.0000	-1.7283	0.0407	20.7401	0

Variable Pair #1:
MALAT1<--CD19
% Change Matrix
Fixed Effect Slope
Coefficient

Variable Pair #2:
FBLN<--CD34
% Change Matrix
Fixed Effect Slope
Coefficient

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-5.0859	-5.5670	143.1615	0.0000
LM-FE	5.3584	0	-0.5069	156.1912	5.3584
LMM-RI	5.8952	0.5095	0	157.4964	5.8952
LMM-RS	-58.8751	-60.9666	-61.1645	0	-58.8751
GEE	0.0000	-5.0859	-5.5670	143.1615	0

Results

Nested Models

Conclusion

Overall Conclusions

Conclusion

Limitations

THANK YOU

References