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 for individual cells
- Hundreds/thousands of SC measurements -- 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
- Multiple-source samples enable analysis of source-level associations
  - e.g., multiple subject sample --> analysis of subject-level associations
- Problematic to integrate multiple samples into a single data set
  - Protocol dependencies affect data quality and reliability

Introduction

Motivation

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

What do I do to solve the problem?

- Outline five modeling methods and how the models account for subject level relationships in SC data
- Apply the modeling methods to motivating SC data example
  - *Describe* how the models account for subject level relationships in the motivating SC data example
  - *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

$(\boldsymbol{X_{ij}}, \boldsymbol{Y_{ij}})$ - subject level predictor-response pair

$\boldsymbol{i = 1, \ldots N}$ - subject from which measurement was taken

$\boldsymbol{N}$ - Total number of subjects

$\boldsymbol{j = 1, \ldots, n_i}$ measurement index taken within subject **i**

(repeated measure index)

$\boldsymbol{n_i}$ - Total number of repeated measurements within subject $\boldsymbol{i}$

Overview of Selected Models

Linear Model (LM)

MODEL DEFINITION:

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 $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

MODEL INFORMATION:

Linear Model (LM)

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

- 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)

MODEL DEFINITION:

Linear Model with Fixed Effect (LM-FE)

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

Terms:

$$(subject)_i = \begin{cases} 0 & if \ \ not \ subject \ i \\ 1 & if \ \ \ \ \ \ subject \ i \end{cases}$$

$$for \ i = 2, \dots, N$$

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

MODEL INFORMATION:

Linear Model with Fixed Effect (LM-FE)

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

- 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)

MODEL DEFINITION:

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

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

Terms:

- β_0: Intercept
- β_1: Fixed Effect Slope
- b_{0i}: Random Effect Intercept (subject) $b_{0i} \sim N(0, \sigma_b^2)$
- $\boldsymbol{\epsilon_i}$: Residual Error $\boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \sigma_\epsilon^2 I_{n_i})$ for $i = 1, \ldots, N$

MODEL INFORMATION:

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

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

- Accounts for subject-level associations by incorporating subject-specific variances
  - Allow for outcomes to be higher/lower for each subject
- Assumes:
  - observations are independent within subject
  - Residual error is independent of the random effects
- $\boldsymbol{\beta_1}$ parameter interpreted as subject-specific representation of relationship between predictor and response
  - Conditional on an individual subject, and
  - Representative of that specific subject

Overview of Selected Models

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

MODEL DEFINITION:

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

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}(subject)_i + \left[b_{1i}(subject)_i X_{ij}\right] + \epsilon_{ij}}$$

Terms:

- β_0: Intercept
- β_1: Fixed Effect Slope
- b_{0i}: Random Effect Intercept (subject)
- b_{1i}: Random Effect Slope(subject)
- $\boldsymbol{\epsilon_i}$: Residual Error $\quad \boldsymbol{\epsilon_i} \sim N(\mathbf{0}, \sigma_\epsilon^2 I_{n_i})$ for $i = 1, \ldots, N$

$$\boldsymbol{b_i} = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N(\mathbf{0}, \boldsymbol{G})$$

$$\boldsymbol{G} = \begin{bmatrix} \sigma_{b_0}^2 & \sigma_{b_{10}} \\ \sigma_{b_{10}} & \sigma_{b_1}^2 \end{bmatrix}$$

MODEL INFORMATION:

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

$$\boldsymbol{Y_{ij} = \beta_0 + \beta_1 X_{ij} + b_{0i}(subject)_i + \left[b_{1i}(subject)_i X_{ij}\right] + \epsilon_{ij}}$$

- Accounts for subject-level associations by incorporating values of:
  - Subject specific, covariate independent variance - allows outcomes to be higher/lower for each subject
  - Subject specific, covariate dependent variance - allows outcomes to be variably associated with covariates according to subject
- Assumes:
  - Residual error is independent of the random effects
- $\boldsymbol{\beta_1}$ parameter interpretation is conditional on an individual subject, and representative of that specific subject

Overview of Selected Models

Generalized Estimating Equations (GEE)

MODEL DEFINITION:

Generalized Estimating Equations (GEE)

Model Requires Individual Specification of:

Used for this Analysis

- Random Component
  - Probability Distribution: assumed for the responses → $Y_{ij} \sim N(\mu, \sigma^2)$
- Systematic Components
  - Linear Predictor: a linear function of the explanatory variables → $\eta_{ij} = \beta_0 + \beta_1 X_{ij}$
  - Link Function: establishes the relationship between Linear Predictor and Expected Outcome → $E[Y_{ij}|X_{ij}] = g(\mu_{ij}) = \mu_{ij} = \eta_{ij}$
- Working Covariance Structure → $[Cov(Y_{ij}, Y_{ik})]_{jk} = \begin{cases} Var(\boldsymbol{Y}_i) & if \ \ j = k \\ 0 & if \ \ j \neq k \end{cases}$

  $j, k \in \{1, \dots, n_i\}$
  - Assumed to approximate true within-subject correlation

MODEL INFORMATION:

Generalized Estimating Equations (GEE)

- Accounts for subject-level association by incorporating assumed correlation structure into residual error
- Iterative fitting process:
  1. Estimate regression parameters (standard GLM Theory), and use to:
  2. Estimate working correlation structure from standardized residuals, and use to:
  3. Correct regression parameter estimates from (1), use to: --> (2)
- Quasi-likelihood, no specification of joint distribution
- A-priori specification of working covariance
  - consistent estimates even with misspecification
- $\boldsymbol{\beta_1}$ parameter interpreted as population average representation of relationship between predictor and response.

Motivating Example

Data

Initial Data:

- Population: 45 Lupus Nephritis Cases vs 25 Control
- Population: 27 subjects, case/control status not present. → 15 Subjects
- 9560 SC observations → 1110 SC Observations
- Over $3.8 * 10^8$ RNA sequencing (scRNA-seq) variable measures → 2 log-transformed Predictor-Response Pairs
- 23 Flow Cytometry variables
- 10 metadata variables (subject, cell-type)

Quality Control Data

Model 1: Predictor log(CD19) --> Response log(MALAT1)
Model 2: Predictor log(CD34) --> Response log(FBLN1)

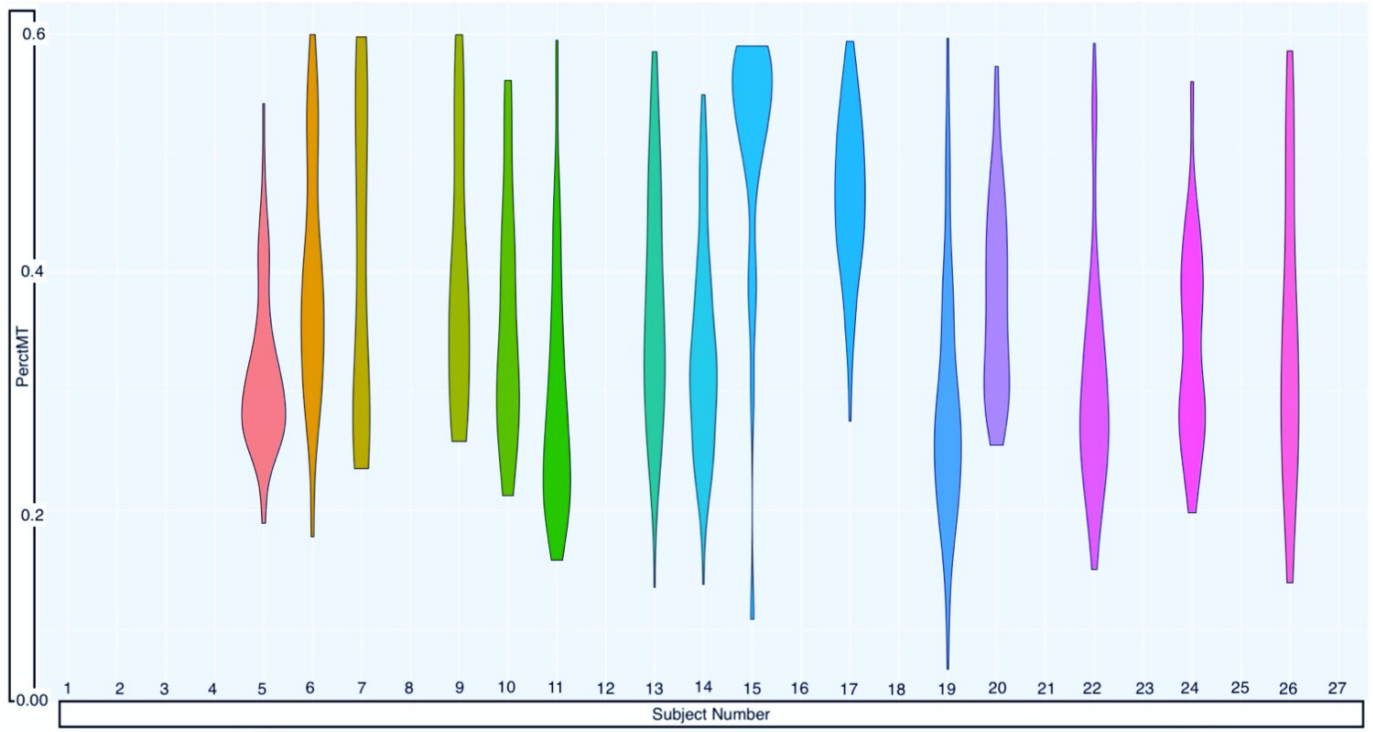
Data Source: 2018 article: “The immune cell landscape in kidneys with Lupus Nephritis patients” [3]

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

Variable Pair #1: MALAT1<--CD19

Variable Pair #2: FBLN1<--CD34

Model Designation	Model Description	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

Fixed Effect Slope Percent Change Matrices

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

Fixed Effect Slope Standard Error Percent Change Matrices

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

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

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

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-30.4878	-30.4878	48.1707	0.0000
LM-FE	43.8596	0	0.0000	113.1579	43.8596
LM-RI	43.8596	0.0000	0	113.1579	43.8596
LM-RS	-32.5103	-53.0864	-53.0864	0	-32.5103
GEE	0.0000	-30.4878	-30.4878	48.1707	0

Fixed Effect Slope Test Statistic Percent Change Matrices

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	3.5197	5.8563	-50.3697	0.0000
LM-FE	-3.4000	0	2.2571	-52.0571	-3.4000
LMM-RI	-5.5323	-2.2073	0	-53.1154	-5.5323
LM-RS	101.4899	108.5816	113.2896	0	101.4899
GEE	0.0000	3.5197	5.8563	-50.3697	0

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

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

Model	LM	LM-FE	LMM-RI	LMM-RS	GEE
LM	0	-4.5977	-1.2994	-41.5292	0.0000
LM-FE	4.8193	0	3.4573	-38.7114	4.8193
LM-RI	1.3165	-3.3418	0	-40.7595	1.3165
LM-RS	71.0256	63.1624	68.8034	0	71.0256
GEE	0.0000	-4.5977	-1.2994	-41.5292	0

Model Parameter Comparison Results

- LM and GEE estimates are similar down to e-4 accuracy
- LM-FE and LMM-RI estimates similar
- LMM-RS Standard Error largest compared to other methods within variable pairings
- LMM-RI Standard Error is smallest compared to other methods within variable pairings
- Test statistics similar between LM and GEE as well as between LM-FE and LMM-RI models
- Test statistics for LMM-RS are 86% larger on average than other models

Results

Nested Models

Nested Model Comparisons:

Testing Inclusion of Fixed Effect Intercept

Variable Pair	Model	Resid DF	RSS	DF	Sum of Squares	F-stat	P(>F)
MALAT1-CD19	LM	1108	1167.76				
	LM-FE	1094	935.89	14	231.87	19.36	6.4776e-44
FBLN1-CD34	LM	1108	650.51				
	LM-FE	1094	214.92	14	435.59	158.38	2.8058e-251

Nested model comparisons: F-test statistics indicating that there is sufficient evidence to support the inclusion of the subject-specific fixed effect intercept

Nested Model Comparisons:

Testing Inclusion of Random Effect Intercept

Variable Pair	Model	df	AIC	logLik	L.Ratio	p-value
MALAT1-CD19	LM	3	3224.097	-1609.048		
	LMM-RI	4	3032.024	-1512.012	194.0722	4.1068e-44
FBLN1-CD34	LM	3	2572.807	-1283.403		
	LMM-RI	4	1438.086	-715.043	1136.72	3.4517e-249

Nested model comparisons: likelihood ratio test statistics indicating that there is sufficient evidence to support the inclusion of the subject-specific random effect intercept

Nested Model Comparisons:

Testing Inclusion of Random Effect Slope

Variable Pair	Model	df	AIC	logLik	L.Ratio	p-value
MALAT1-CD19	LMM-RI	4	3032.024	-1512.012		
	LMM-RS	6	2993.820	-1490.910	42.20503	6.8437e-10
FBLN1-CD34	LMM-RI	4	1438.086	-715.043		
	LMM-RS	6	1438.068	-713.034	4.018095	0.1341

Nested model comparisons: likelihood ratio test statistics indicating that there *is* sufficient evidence to support the inclusion of the subject-specific random effect slope into the LMM-RI model for the MALAT1 ~ CD19 variable pairing.

Conclusion

Overall Conclusions

- Population average models (LM and GEE) & models with subject-specific intercept terms (LM-FE and LMM-RI) have:
  - Similar estimates within/differing estimates between descriptions
- Nested Model Comparisons: sufficient evidence to support inclusion of:
  - Subject specific fixed effect intercept
  - Subject specific random effect intercept

Indicative of subject-specific, covariate independent associations not accounted for in overall population averaged/marginal models

- LMM-RS model has largest standard errors
- Nested model comparisons: *borderline* or insufficient evidence to support inclusion of subject-specific random slope

Indicative that subject-specific associations are NOT covariate dependent

Conclusion

Limitations & Future Work

Limitations

- All results based on just two scRNA-seq variable pairs
- scRNA-seq data heavily influenced by protocol dependencies & measurement inconsistencies

Future Work

- Extending analysis to all variable pairs
- Overfitting considerations: test/train model development not implemented but should be going forward

THANK YOU

References

1. Bacher R, Kendziorski C (2016) Design and computational analysis of single-cell rna-sequencing experiments. *Genome biology* 17: 63.

2. Ståhlberg A, Andersson D, Aurelius J, et al. (2010) Defining cell populations with single-cell gene expression profiling: Correlations and identification of astrocyte subpopulations. *Nucleic acids research* 39: e24–e24.

3. Arazi A, Rao DA, Berthier CC, et al. (2018) The immune cell landscape in kidneys of lupus nephritis patients. *bioRxiv* 363051.