Comparing Correlated Data Models on Single-Cell RNA Expression Profiles

Lee Panter

Monday, November 25, 2019

Presentation Overview

Project Goals and Desired Outcomes:

- Develop multiple statistical models for Single-Cell RNA Sequencing data
- Compare the models for: fit, estimate stability, and diagnostic integrity.
- Suggest a model.

Presentation Highlights:

- Introduction to RNA and Single-Cell
- Data Summaries and Proposed Modeling Approaches
- Results, Comparisons, and Conclusions
- Future Research, Outstanding Problems, Areas of Interest

Introduction to RNA Sequencing

RNA Sequencing (RNAseq) [1]

- Which genes are being expressed and at what magnitude?
- How do gene expressions change over time, or between treatment groups?
- Used in:
  - Transcriptional Profiling
  - Single Nucleotide Polymorphism (SNP) identification
  - Differential Expression

RNAseq Expression Profiles

- Count data – higher values $\Rightarrow$ higher level of expression
- Genes $\rightarrow$ (on/off)? $\Rightarrow$ Expression Value is $(0 \quad or \quad > 0)$
- Indicative of zero-inflation

Single-Cell Methods

Single-Cell (sc) Data:

- Measurements single-cell resolution
- *Batch-Samples* from subjects $\Rightarrow$ Single-Cells "sub-sampled" from each $=$ Observational Units.

Repeated Measure/Clustering Assumptions:

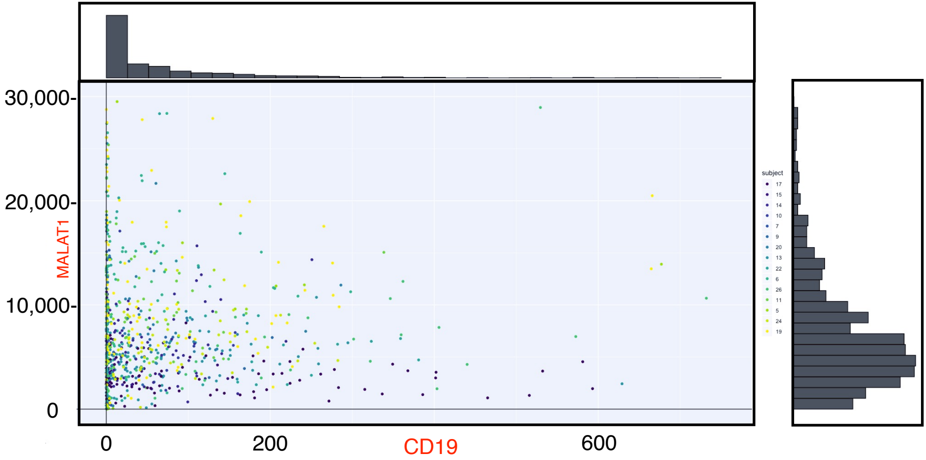
- SC observations are independent between Batch Samples
- Covariance between all Batch Samples assumed to be identical

Case Study scRNA-seq Data:

- $\sim 38 * 10^3$ variables (genes), $\sim 9 * 10^3$ observations (SCs) [2]
- Poor measurement accuracy. Problems with: batch effects, contamination, duplicate reads,...etc. [3]
- Quality control filtering: $\sim 9 * 10^3$ obs $\longrightarrow \sim 1,000$ obs

scRNA-seq Data Summary

MALAT1 vs CD19



NOTE 223 extreme observations removed to enlarge main distribution

Proposed Modeling Approaches: Notation – OLS & LMM

Notation:

- Fixed Effects:
  - Global Intercept:

    $\sim 1 + \cdots$
  - Subject Factor:

    $\sim subject + \cdots$
  - Covariate Factor:

    $\sim CD19 + \cdots$
- Random Effects:
  - Intercept:

    $\sim (1|subject) + \cdots$
  - Slope:
    $\sim (CD19|subject) + \cdots$

OLS and Linear Mixed Effects Models

- OLS:
  - Predictors:
    $\sim$1 + CD19
- LMM:
  - Fixed Effects:
    $\sim$ 1 + CD19
  - Random Effects:
    $\sim$ (1 | subject)

    + (0 + CD10 l subject)
  - Repeated Measures:
    Unstructured (CS)

Proposed Modeling Approaches: Generalized Linear (Mixed) Models

- Poisson Regression (No Over-dispersion) & Poisson Quasi-Likelihood (w/Over-dispersion)
  - Error Distribution: Poisson
  - Linear Predictor: 1 + CD19
  - Link Function: log
- Generalized Linear Mixed Models (Penalized QL) [4]
  - Error Distribution: Poisson
  - Linear Predictor(s):
    FIXED=1 + CD19
    RANDOM=(1 | subject) + (0 + CD19 | subject)
  - Link Function: log

Proposed Modeling Approaches: Zero Inflated Poisson [5]

Occurrence Model: $\boxed{R_{ij} \sim bernoulli\left(p_{ij} | a_0, a_1\right)}$
where a_0, a_1 are Occurrence-Model random effect parameters

Intensity Model: $\boxed{Y_{ij} | \left(r_{ij} = 1, a_0, a_1\right), \sim Poisson\left(\lambda_{ij} | b_0, b_1\right)}$
where b_0, b_1 are Intensity-Model random effect parameters

Zero-Inflated Poisson, Generalized Linear (Mixed) Models

Fit Using Adaptive Gauss-Hermite Quadrature

- Error Distribution: "Zero-Inflated Poisson"
- Occurence & Intensity Model Linear Predictors:
  - Fixed Effects: $\{\sim 1, \sim 1 + CD19\}$
  - Random Effects: $\{\sim 1, \sim 1 + CD19\}$
- Link Function: Log

Results, Comparisons, Conclusions

Model	Intercept Estimate	Std.Err	p-value
LMwFE	$7.7624 * 10^3$	$2.3480 * 10^2$	$< 2 * 10^{-16}$
LMMwRE	$7.338 * 10^3$	$7.6776 * 10^2$	$< 2 * 10^{-16}$
POI	8.957	$3.723 * 10^{-4}$	$< 2 * 10^{-16}$
POIql	8.957	$3.007 * 10^{-2}$	$< 2 * 10^{-16}$
POIqlLMM	8.8362	$1.0160 * 10^{-1}$	$1.7 * 10^{-3}$
ZIP	8.9572	$< 2 * 10^{-4}$	$< 2 * 10^{-4}$

Model	Slope Estimate	Std.Err	p-value
LMwFE	$7.1320 * 10^{-1}$	1.5426	$6.440 * 10^{-1}$
LMMwRE	2.168	1.797	$2.278 * 10^{-1}$
POI	$8.839 * 10^{-5}$	$2.369 * 10^{-6}$	$< 2 * 10^{-16}$
POIql	$8.839 * 10^{-5}$	$1.913 * 10^{-4}$	$6.440 * 10^{-1}$
POIqlLMM	$3.16 * 10^{-4}$	$1.653 * 10^{-4}$	$5.61 * 10^{-2}$
ZIP	$1 * 10^{-4}$	$2.03 * 10^{-6}$	$< 2 * 10^{-16}$

Note: $e^{8.957} \approx 7.762 * 10^3$

Results, Comparisons, Conclusions

Conclusions Drawn from Results:

- Simpler models performed better according to the AIC criterion
- Parameter estimates for global intercept showed higher stability and significance than estimates for slope

Model	AIC
LMwFE	$2.2851 * 10^4$
LMMwRE	$2.2851 * 10^4$
POI	$5.7046 * 10^6$
POIql	NA
POIqlLMM	NA
ZIP	$4.1791 * 10^6$

Future Research, Outstanding Problems, Areas of Interest

Outstanding Issues:

- Comparing quasi-likelihood models to linear models and quadrature methods

Future Research & Areas of Interest:

- Log-transformed responses, additional variable combinations, marginal average models

Thanks for Listening!

If You Want To Learn More:

- email: lee.panter@ucdenver.edu

- Project GitHub: https://github.com/leepanter/BIOS6643FinalProject.git