Project Write-Up Outline

Title: Comparing Models over Single-Cell Data

Goals of paper

- Establish a need for developing single-cell data modeling methodologies
- Propose new candidate modeling methods
    - Why are the new methods superior to existing methods?
    - What are the theoretical advantages of the new methods?
    - What are some possible downsides of the methods?
- Establish a basic theoretical understanding of new modeling approaches
    - What changes from existing approaches?
    - What stays the same?
    - Use theory to establish expected results
- Apply new approaches to case study data and compare outcomes to hypothesized results.
- Propose alternative modeling approach based upon results of case study analysis (and other factors as needed).

Abstract:
Single-cell RNA sequencing data collected to improve estimates of RNA expression variability has traditionally been modeled using unsupervised methods such as differential expression testing and gating. These methods may be easily implemented, but they ignore observational clustering inherent to single-cell data. This paper looks to compare three different modeling strategies, and multiple models within each strategy that account for the correlated nature of single-cell data. The modeling approaches will be compared theoretically against a standard Ordinary Least Squares model. Additionally, single-cell data from a Lupus Nephritis case study will be used to compare the models analytically. It is hoped that this paper will present new approaches to modeling single-cell data, and will be useful for not only Statisticians, but also Geneticists and Microbiologists.

Introduction

Section Goals:

- Describe paper's primary motivation and objective
- Process used to accomplish objective: a brief outline of paper's structure

Outline:

- What is single-cell RNA sequencing (scRNAseq) Data?

    - Traditional RNA sequencing methods used recombination methods that only allowed for estimates of parameters for population averages
    - Single-cell RNA sequencing allows for estimation of cell-cell variability
    - Can be useful for targeting cancer & lupus treatments, among others
- How is single-cell data traditionally modeled?
    - Traditional modeling techniques include:
        - Unsupervised clustering & gating
        - Traditional modeling approaches such as OLS
    - Traditional methods fail to account for nested observational structure
        - Single-cells each taken as observational units
        - Many single-cell OUs sourced from the same sample
    - New Modeling Methodology needed that can accurately and precisely model single-cell data while accounting for correlation of repeated measures within sample
- Basic outline of this paper
    - Proposal of modeling methodologies & theoretical comparisons
    - Case Study Analysis
    - Discussion
    - Future Research

Proposal of Modeling Methodologies & Theoretical Comparisons

Section Goals:

- Give enough detail on each model to fully define
- Provide relevant differential theory between OLS and subsequent modeling methods
- Use theory to hypothesize effect of methodological alterations on model estimates, standard errors,…etc.
- Motivate usage of model theoretically, based upon motivation given in introduction/abstract.

Outline

- Linear Models: no theoretical details, just parameters, and definition of models
- Linear Mixed Effects Models:
    - What's different from OLS?
    - How/Why do these models account for observational clustering?
    - How do these differences theoretically impact the resulting model when fitting clustered data?
- Generalized Estimating Equations
    - How is this different from the other two methods?
    - Why do these models account for observational clustering?
    - How do these differences theoretically impact the resulting model when fitting clustered data? And how should this compare to Linear Mixed Effects Models?

- How are outcomes compared, and judged to be better/worse? (model selection criteria)
    - AIC of model
    - Test set MSE

Case Study Analysis

Section Goals:

- Provide appropriate result information to support deductions that will support conclusions in discussion
- Provide functional background on Lupus paper
- Provide Results of Model Fits

Outline:

- Information on the Data
    - "The Immune Cell Landscape in kidneys of Lupus Nephritis Patients"
    - Citation, link to data
    - Study design, dimension of data, data type
    - Quality control metrics used
    - Variables selected for test regressions & motivation for their selection
    - Initial data summaries
        - Histograms-predictors & outcomes
        - Scatter plots
    - Transformations used & transformed data summaries
- Results
    - Tables for Intercept and Slope
        - Estimate, standard error, test statistic, p-value
    - Percent change matrices for intercept and slope:
        - Estimate
        - Standard error
    - Plots of:
        - Residual V Fitted Values
        - Model V Original Data
        - QQ Plot

Discussion

Section Goals:

- Determine if results are in agreement with hypothesis
    - Evaluate possible issues if conflicting

    - Demonstrate agreement using specific results
- Address questions outlined in “Goals of Paper” section using information from either
    - Proposal of modeling methodologies & theoretical comparisons
    - Case Study Analysis
- Address assumptions, reductions, possible errors

Outline:

- Hypothesis V Results
    - If conflicting
        - Cite specific results and theory to demonstrate conflict
        - Evaluate possible sources of error in analysis
            - Specifically consider Type II sources of error
        - Make a recommendation for rectifying
    - If agreeing
        - Cite specific results and theory to demonstrate agreement
        - Evaluate possible faults in analysis
            - Specifically consider Type I sources of error
- Questions from Goals of Paper
    - Why are the new methods superior to existing methods?
        - Use: specific citations from clustering theory, and results to make relevant claims
    - What are the theoretical advantages of the new methods?
        - In addition to above answer
            - Empirical Bayes estimation?
    - What are some possible downsides of the methods?
        - Complexity
        - More to come

Future Research

Section Goals:

- Enumerate unanswered, or partially answered questions
- Propose new research that might answer those questions

Outline

- Issues still outstanding
    - What is the issue?
    - Why couldn’t the issue be resolved?
    - Can the issue be resolved using the current information?
        - If not, what information is needed to resolve it?
- If new information is still needed, is it:

- More data?
- More/conflicting theory?
- A new approach?