Comparing Models of Subject-Clustered Single-Cell Data

Lee Panter

Audrey Hendricks\*, PhD-Committee Chair and Advisor

Stephanie Santorico[[1]](#footnote-1), PhD-Committee Member

Rhonda Bacher[[2]](#footnote-2), PhD-Committee Member

# Abstract

# Introduction

Traditional methods of sequencing the human transcriptome involve analyzing the combined genetic material of thousands or even millions of cells. These so called “bulk” techniques provide information about the average gene expression across the cell sample but often fail to capture the underlying variability in expression profiles within the sample of cells [1].

Conversely, single-cell RNA sequencing (scRNA-seq) obtains measurements of transcriptomic information from hundreds or even thousands of RNA-sequencing profile measurements, each specific to a single-cell. This feature of single-cell data analysis is suited for research applications that seek to identify rare cellular subpopulations or characterize expressions that are differentially expressed across conditions [2]. Additionally, technological developments have made generating single-cell data more cost effective, and easier to obtain on multiple sample-sources, most notably on multiple individuals.

The utility of single-cell data, and the feasibility of single-cell data measurements across multiple subjects motivates a need to compare methods that can adequately model single-cell data with subject level associations attributable to observation sampling nested within subjects.

Here, I compare five methods for modeling scRNA-seq expression profiles that account for within-subject correlation: linear modeling (LM), linear modeling with subjects as fixed effects (LM-FE), linear mixed effects models with subjects only as random intercepts (LMM-RI) or as both random intercepts and random slopes (LMM-RS), and generalized estimating equations (GEE). I first present the overall framework for each method. Then I compare the results for each model using single-cell data from a study of 27 Lupus Nephritis cases.

# Description of Data Set

Throughout this paper references are made to the 2018 article entitled “The immune cell landscape in kidneys with lupus nephritis patients”, in which Arazi, Rao, Berthier, et al. compare single-cell kidney tissue sample data from 45 Lupus Nephritis subjects vs. 25 population controls [3]. The kidney tissue samples were collected from ten clinical sites across the United States, cryogenically frozen, then shipped to a central processing facility. At the central processing facility, the tissue samples were then thawed, and sorted into single-cell suspension across 384-well plates using FlowJo 10.0.7, 11-color flow cytometry [4]. Single-cell RNA sequencing was performed using a modified CEL-Seq2 method [5] with 1 million paired-end reads per cell. The original experimental data may be accessed by visiting the Immport repository with accession code SDY997. [Immport-SDY997: https://www.immport.org/shared/study/SDY997](https://www.immport.org/shared/study/SDY997)

The original research conducted in Arazi, Rao, Berthier, et al, concerned 70 subjects. A subset of the original research data was made available for the purposes of this analysis that contained 9,560 single-cell observations originating from 27 subjects.

In each single-cell observation there are:

* Over RNA sequencing measurements
* 23 Flow Cytometry measurements
* 10 meta data-variables (e.g. subject of origin, cell type)

I implement a quality control (QC) process to filter observations that are inadequately representative of living, single-cell, samples from an agglomerated Lupus Nephritis case/control population. Appropriate quality control filter thresholds are chosen from calculations involving all RNA sequencing measurements. Observations within each subject classified inadequate (poor/low quality) are filtered out if they are either: not a single-cell (i.e. multiple, partial, or missing cells), or cellular material that is insufficiently alive. After quality control filters are imposed, 1110 observations originating from 15 subjects remain for analysis. Details related to the QC filtering process are contained in .

I focus on two pairs of RNA sequencing measurements for model comparisons. I use the two outcome variables motivated by previously established associations with human disease conditions [6] [7]. I calculate pairwise correlations for each of the chosen outcome variables, and I choose the Cluster of Differentiation marker (see for description) with the highest correlation to pair as a predictor to each of the outcome variables. I perform a log-transformation on the predictor and response of both variable pairings motivated by each variable’s right-skewed distribution.

I use the following transformed variable pairings to perform model comparisons:

Further details related to variable selection and variable summaries are contained in .

# Model Descriptions

In the following sections a description is provided for each model using the following notation for a subject level predictor-response pair:

where represents the observation’s (subject from which the measurement was taken), and represents the measurement index taken within subject (the repeated measure index within each subject).

## Linear Model (LM)

The linear model can be written as:

This model does not account for correlation structure in the data, and instead assumes the observations are independent. Linear model parameter estimates are for the population averages.

The error term, , is assumed to be a normally distributed random variable with mean zero and variance .

## Linear Model with Fixed-Effect (LM-FE)

Adding a subject specific fixed effect intercept term to the LM model allows for the accounting of subject level effects by uniformly shifting the mean of the fitted values specific to a subject. This model is written as:

where

$$ \mathbb{I}\left(subject\_{i} \right)= \begin{cases} 1 & \mbox{if} \quad subject=i \\ 0 & \mbox{if} \quad subject \neq i \\ \end{cases} \quad \text{for} \quad i=2,\ldots,N\\ $$

This model adds estimated parameters which represent the average deviation of each subject from the global estimated mean Linear Model (LM).

## Linear Mixed Effects Models

Linear mixed effects models that incorporate multiple subjects using random effects are the next methods outlined. Linear mixed effects models do not require the assumption of independent observations. Structures such as autoregressive, moving-average, or simply unrestricted (unstructured) can be used to explicitly model within-subject correlation. Additionally, random effects can be incorporated and fit with covariance parameters that capture between-subject effects.

### Linear Mixed Effects Model with Random Intercept (LMM-RI)

A linear mixed effects model with a random intercept controls for subject-level correlations through the use of subject specific variances. The LMM-RI model is written as:

where

it is assumed that and are independent.

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

A random slope linear mixed effects model differs from each of the previously considered methods because it allows for distinct relationships for each subject between the predictor and response variables of interest. The LMM-RS model is written as:

where for each subject

is the identity matrix, and it is assumed that and are independent.

## Generalized Estimating Equations (GEE)

The final modeling method considered is generalized estimating equations (GEE). The GEE framework requires the specification of a systematic and random component. It also requires the specification of an assumed covariance structure which approximates within-subject correlation, and which the GEE algorithm iteratively re-fits estimated values. Each iteration of the GEE algorithm incorporates information about all subjects into successive estimates of parameters.

The components for the GEE model are:

* The random component
  + A probability distribution is assumed for the responses. The normal distribution is assumed here.
* The systematic component
  + The linear predictor, , is a linear combination of the predictors. Here, there is only one predictor (), and the linear predictor used is:
* The link function
  + The link function provides the relationship between the linear predictor and the expected outcome, i.e:
* Working Covariance Structure
  + An independent working covariance structure is used here:
  + for each subject

Estimates for GEE parameters are calculated by solving an using the Newton-Raphson iterative root-finding algorithm. Detailed method descriptions, including derivation and solving of the estimating equations can be found in Fitzmaurice, Laird, and Ware [8].

The GEE algorithm is robust to misspecification of the working covariance structure. This means that initially incorrect specifications of the working covariance matrix still converge to the appropriate structure. This stability is due in-part to the fact that the method estimates population average effects. This stability is also attributable to the fact that GEE models the relationship between response and covariate separate from an initially assumed, then iteratively recalibrated correlation structure of the repeated measures within grouping [9].

## Parameter Interpretations

The LM and GEE modeling methods are techniques used for obtaining estimates of population averaged fixed effect slope parameters. These parameter values are interpreted as contributing to the response of the average subject (not representative of any single subject within the sample). An example interpretation of this parameter is: **across all subjects, a one-unit increase in the predictor () is associated with a unit change, on average, outcome () (assuming all other covariates are held constant).**

The LMM-RI and LMM-RS modeling methods are techniques used for obtaining estimates of subject specific fixed effect slope parameters. These parameter values are interpreted as effects conditional on a specific subject, contributing to the response of the specific subject. An example interpretation of this parameter is: **after having conditioned on the effects of a specific subject, a one-unit increase in the predictor () is associated with a unit change in the expected outcome () of that same subject (assuming all other covariates are held constant).**

Finally, the LM-FE method is a technique used for obtaining estimates of population averaged fixed effect slope parameters adjusting for average subject effects. These parameters are interpreted as contributing to the response of the average subject after adjusting for a each subject’s average effect. An example interpretation of this parameter is: **across all subjects, after having adjusted for the average effect of each subject, a one-unit increase in the predictor () is associated with a unit change, on average, in the outcome () (assuming all other covariates are held constant).**

# Results

## Parameter Value Comparisons

A comparison of main effect slope coefficient, standard error and test statistic (shown in below) across modeling approaches within variable pairings indicates that estimates produced by the LM and GEE methods are similar down to . The LM-FE and LMM-RI method estimates are also similar since estimates for each parameter type (coefficient, standard error and test statistic) exhibit magnitude and directional similarities in both variable pairings.

Displayed in - below are percent change in: parameter estimate, standard error, and test statistic for the variable pairing in and the variable paring in . Where the percent change is defined as:

reinforce that the LM and GEE model estimates are comparable within variable pairings. A similar conclusion conclusion can be drawn for comparisons of the LM-FE and LMM-RI model.

The percent change tables also show that the LMM-RS estimates for the fixed effect slope parameter standard error is the largest when compared to the corresponding estimates within variable pairing as generated by other modeling methods. In contrast, the standard error of the fixed effect slope parameter is smallest for the LMM-RI model within variable pairings.

The differences in test statistics of the fixed effect slope parameter for each modeling method within each variable pairing are analogous to the differences in slope coefficients previously noted. In particular, test statistics have similar values between the LM and GEE models as well as between the LM-FE and LMM-RI models. Test statistics calculated for the LMM-RS model are the least similar to the other modeling methods. The LMM-RS model averages test statistics that are 86% larger than the other models. This decreased similarity is also accompanied by decreased parameter significance.

## Nested Model Comparisons

() above is a nested model comparison, the result of which is an F-test statistic indicating that there is sufficient evidence to support the inclusion of the subject specific fixed-effect intercept into the LM model.

() above is a nested model comparison, the result of which is a likelihood ratio statistic indicating that there is sufficient evidence to support the inclusion of the subject specific random effect intercept into the LM model.

() above is a nested model comparison, the result of which is a likelihood ratio statistic indicating that there is sufficient evidence to support the inclusion of the subject specific random effect slope into the LMM-RI model for the variable pairing. However, there is insufficient evidence to support the inclusion of the subject specific random effect slope into the LMM-RI model for the variable pairing.

# Discussion

Here, I compared five modeling strategies for detecting subject level associations in single-cell RNA sequencing data gathered over 27 subjects from a Lupus Nephritis study: linear modeling (LM), linear modeling with subjects modeled as fixed effects (LM-FE), linear mixed effects models with subjects modeled as only random intercepts (LMM-RI) or random intercepts and random slopes (LMM-RS), and generalized estimating equations (GEE).

Each of the stated results is representative of some type of subject level association within the single-cell RNA sequencing data I investigated. The noted differences between estimates produced by population average interpretation models LM/GEE and those produced by the subject specific interpretation model LMM-RI is indicative of subject specific, covariate independent associations not explicitly modeled by the overall population averaged model. Similarly, the fixed effect parameter differences noted in comparisons with the LM-FE model estimates are indicative of population average, covariate independent associations not explicitly modeled in the comparative model. Finally, an estimate that differs from those generated by the LMM-RS method are indicative of a subject specific, covariate dependent association not otherwise accounted for by the other modeling technique.

In conjunction with the information gained through parameter estimate comparisons, the nested model comparisons allow for further inference on specific types of subject level associations. There is evidence for the inclusion of all covariate terms into all models except for the random slope into the LMM-RI model in the case of the variable pairing. This coincides with intuition as differences were noted between estimates generated by LMM/GEE compared to LM-FE/LMM-RI in both variable pairings. The LMM-RS model was noted as having the largest standard error, in addition to having the least similar estimate values. These findings illustrate that there is sufficient evidence for subject-level associations that are covariate independent, but that there is insufficient or questionable evidence to support any subject-level associations that are covariate-dependent.

The analysis I conducted in this paper has detected a variety of subject level associations in single-cell RNA sequencing data using model comparison techniques. I have detected subject level associations that are related to both covariate dependency and parameter interpretation (population average or subject specific). I also used nested model comparisons to determine the strength of evidence of the subject level associations I detected.

The analyses performed here are subject to several drawbacks and limitations. All the results are based on evidence obtained from just two single-cell RNA sequencing variable pairings. In the future, comparing the consistency of these models over all model pairs is needed. Additionally, single-cell RNA sequencing data is heavily influenced by protocol dependencies and measurement inconsistencies. Quality control must be carefully considered and conducted prior to any analysis.

The utility and promise of single-cell RNA sequencing data indicates that such data will become more prevalent and will be extended to multiple subject samples. I have presented an initial comparison of methods for detecting subject-level associations in single-cell RNA sequencing data sets.

# Appendix

## Appendix A: Data Quality Control

I use the Seurat Guided Clustering Tutorial [10] to perform quality control (QC) of the initial data. This process quantifies the quality of each single-cell observation in two numerical measures (based upon two calculated variables, and ). Threshold values of these variables are chosen and used to filter cells (observations) not meeting the chosen criteria. The Seurat tutorial provides methods of automated calculation and filtering implemented by Arazi, Rao, Berthier, et al. in [3]. Identical variable calculations, with alternative threshold settings are independently implemented for this study.

The quality control variables are conceptually defined as:

1. is the number of unique genes detected to have a non-zero expression in each cell. This is used to identify cells with an abnormally low or high number of expressed genes. Low numbers may result from empty wells (zero content measurements) or broken (partial) cells, while high numbers may result from observations of more than one cell.
2. is the percentage of reads that map to the mitochondrial genome. This is used to identify dead and/or broken cells as dead or dying cells will retain RNAs in mitochondria, but lose cytoplasmic RNA [2].

The pre-QC distribution of for each subject is displayed in () below:

The QC measure thresholds employed by Arazi, Rao, Berthier, et al. in [3] are:

All observations for which the calculated values of and satisfied the inequalities in (1) and (2) above were kept, and the others were considered “low-quality” and removed. The resulting distribution of the variable is displayed in ():

As 84% of cells as removed with the filters chosen by Arazi et al, I choose a more lenient threshold, removing observations with , in an effort to keep more cells.

An additional restriction of the data to only B-cells is made in an effort to regularize the data sample (i.e. homogenize feature expression). The resulting distribution of is displayed in () after filtering.

The distribution of observations for each subject Pre/Post QC (with updated threshold value) is shown numerically in ():

## Appendix B: Variable Selection and Summaries

I select two pairs of variables from the 38,354 genetic markers in the Lupus Data to compare across the five modeling methods. The variables I choose have higher values of correlation than arbitrary variable pairings as indicated by a high Pearson Correlation Coefficient (both selected pairings are within the top 10% of highest Pearson Correlation Coefficients of all possible pairings), and have previously been associated with human diseases or conditions (e.g. cancer treatment research in the case of MALAT1 [6]-used as the first outcome, or observed limb malformations in the case of FBLN1 [7]-used as the second outcome). I also attempt to assign predictor-pairings of interest. The CD19 marker (the predictor paired with MALAT1) is a transmembrane protein encoded by the CD19 gene. The FlowJo cytometry measurements contain CD19 protein readings, so the relationship between CD19 as a predictor and the outcome of interest (MALAT1) can be modeled using proteomic or transcriptomics data. CD34, the predictor which I link with FBLN1 is also a transmembrane protein encoded by a gene, and similarly interesting.

Without undergoing the process of expression normalization, single-cell RNA sequencing data is represented as non-negative integer count values. Higher counts correspond to higher detection frequencies and these detection frequencies can be interpreted as a quantification of the magnitude of expression for a transcriptomic marker (e.g CD19, CD34, MALAT1, FBLN1).

I provide numerical summaries of the four selected variables in . Each describes selected variable summary statistics (minimum, maximum, average, and median) for the positive observational count subjects in ().

Measurements of scRNA-seq data are specific to precise transcriptomic targets. This means that single-cell expression profiles (a single observation) can be limited to a small transcriptomic scope. So while the agglomerated scope of gene expression across a sample is the same as (or broader than) a traditional bulk experiment, individual observations have a biologically inflated zero-component. There are also zero-inflation components that are associated with protocol variations, and measurement error. Together, these factors contribute to right-skewed variable distributions ()

The MALAT1 variable has a large minimum outcome compared to the other variables, so I translate all the values of this variable by the minimum value.

This gives a minimum expression value of zero, which coincides with intuition as well as the minimum value of the other variables under investigation.

The modeling methodologies I employ motivate a log-transformation in an attempt to achieve approximate normality, especially for the outcome variable’s distribution. I perform the “log plus +1” transformation on all variables (predictor and outcome):

The resulting distributions are shown in ()

The log-transformed response of MALAT1 is approximately normally distributed; however, the log-transformed response FBLN1 is not inherently better than the un-transformed response.

Regardless, each outcome is modeled under the assumption that: compensating for observational correlation will sufficiently account for non-normality of the responses. It may be the case that additional transformations and/or alternative modeling techniques may be needed to improve model error distributions. However, for the purpose of comparing the previously mentioned models on subject-correlated single-cell data, I proceed with this assumption and I verify residual homoscedasticity, normality and independence using fitted vs residual plots and quantile-quantile plots.

### Variable Summary Tables

are summary tables for the variables chosen in this analysis. The data in each table pertains to a subject that had non-zero post quality control observation counts (i.e. the subject had data that past quality control filters). All values displayed are calculated on post-qc data.

# Code and Data

All code for the above analysis was written and evaluated in RStudio Version 1.2.1335, and is available for download at the following GitHub repository:

<https://github.com/leepanter/MSproject_RBC.git>

Additionally, a link to all necessary and reference data files (including original data) are contained in the following Google Drive:

<https://drive.google.com/open?id=1gjHaMJG0Y_kPYWj5bIE4gRJU5z9R2Wqb>

# References

1. Macaulay IC, Voet T (2014) Single cell genomics: Advances and future perspectives. *PLoS genetics* 10: e1004126.

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

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

4. FlowJo X V10. 0.7 r2 flowjo. *LLC https://www flowjo com*.

5. Hashimshony T, Senderovich N, Avital G, et al. (2016) CEL-seq2: Sensitive highly-multiplexed single-cell rna-seq. *Genome biology* 17: 77.

6. Gutschner T, Hämmerle M, Diederichs S (2013) MALAT1—a paradigm for long noncoding rna function in cancer. *Journal of molecular medicine* 91: 791–801.

7. Debeer P, Schoenmakers E, Twal W, et al. (2002) The fibulin-1 gene (fbln1) is disrupted in at (12; 22) associated with a complex type of synpolydactyly. *Journal of medical genetics* 39: 98–104.

8. Fitzmaurice GM, Laird NM, Ware JH (2012) Applied longitudinal analysis, John Wiley & Sons.

9. Odueyungbo A, Browne D, Akhtar-Danesh N, et al. (2008) Comparison of generalized estimating equations and quadratic inference functions using data from the national longitudinal survey of children and youth (nlscy) database. *BMC medical research methodology* 8: 28.

10. Satija R, others (2018) Seurat: Guided clustering tutorial. *Satija Lab http://satijalab org/seurat/pbmc3k\_tutorial html*.

1. The University of Colorado-Denver [↑](#footnote-ref-1)
2. The University of Florida [↑](#footnote-ref-2)