Comparing Models of Subject-Clustered Single-Cell Data

Version 2.0

Lee Panter

Abstract

Single-cell RNA sequencing (scRNA-seq) represents a revolutionary shift to the analytic approaches being used to decode the human transcriptome. Single-cell is used to: visualize cellular subpopulations with unsupervised clustering methods, test for differential expression rates across conditions using logistic and mixture modeling, and reconstruct spatio-temporal relationships in the microbiome using network analysis. These accomplishments demonstrate the utility and promise of single-cell research; however, if numerical results are desired, each analysis needs to be altered upon the hypothetical inclusion of single-cell observations sourced from multiple individuals. Since single-cell data acquisition is increasing in efficiency and decreasing in cost, data sets featuring single-cell observations from multiple subject sources can be expected to rise in prevalence as a default method of attempting to improve analytic power. Therefore, there is a practical need to outline, analyze, and compare current methods for obtaining numerical parameter estimates for between-subject observation correlation. This paper looks to compare three different modeling strategies (each with different estimates for between-subject correlation parameters) for scRNA-seq expression estimation in data with subject-level clusting. The modeling approaches are compared theoretically, and analytically, motivated by data from a Lupus Nephritis study. It is hoped that this paper presents insights

into modeling single-cell expression data, as well as aids researchers with down-stream analyses, and future theoretical/analytic methodology development.

Introduction

Single-cell analysis has emerged as a leading methodology for transcriptome analytics. [1] Single-cell data sets (i.e. data involving measurements with single-cell resolution) demonstrate their utility in research contexts for identifying rare subpopulations, characterizing genes that are differentially expressed across conditions, and infering spatio-temporal relationships within the microbiome. [2] Additionally, advances in whole genome amplification and cellular isolation techniques make single-cell data sets more accessible, more informative, and more diverse than ever before. [1] Therefore, there is a clear need to compare, test, and integrate methods that can accurately and precisely model single-cell data and account for the correlation of repeated measures within subject samples.

This paper seeks to satisfy this need by comparing three methods for modeling scRNA-seq expression profiles that account for within-subject correlation differently. We compare theses parameter estimates obtained using data consisting of scRNA-seq observations across multiple subjects with Lupus Neprhitis. General modeling theory is provided in the context of this example and we discuss relevant conclusions, implications, limitations and future research to illustrate our findings.

Previous Results

The following studies use single-cell data to make “down-stream” conclusions. A down-stream analysis will incorperate information generated from a statistical study to make conclusions about relateable biological concepts. During this process, the conclusions drawn from statistical inference are logically equated to biological implications. Therefore, each "down-stream" result is dependent upon a coherent statistical analysis. The examples below show that coherent statistical inference will be unreasonable when the underlying data exhibits subject-clustering.

Sub-Population Detection

Traditional methods for subpopulation exploration within single-cell data commonly involve unsupervised clustering techniques including Principle Components Analysis (PCA) and K-Nearest Neighbors (KNN). These methods can effectively identify rare nerological cells within a homogeneous population. [3] Such clustering methods, and additional (non-linear) methods such as the t-distributed stochastic neighborhood embedding (t-SNE) are also useful for visualizing high-dimensional data are used to find multi-dimensional boundary values for distinguishing heathly and cancerous bone marrow samples. [4] While all these studies involve single-cell data that incorperates multiple subjects, the modeling methodologies do not provide numerical estimates for the effects of subject-clustered sampling, and therefore can only be used heuristically.

Test for Differential Expression Across Conditions

Single-cell data is used to target treatments by characterizing differential expression across condition. Model-based Analysis of Single-cell Transcriptomics (MAST) is used to compare "primary human non-stimulated" and "cytokine-activated" mucosal-associated invariant T-cells. [5] Additionally, Single-Cell Differential Expression (SCDE) is used to compare 92 embryonic mouse fibroblasts to 92 embryonic human stem cells. [6] Neither of these studies included samples across multiple subjects (excluding paired/treatment sample assumptions used for parameteric tests).

Investigate Spatio-Temporal Microbiome Relationships

Network modeling approaches, in conjunction with single-cell data provides the opportunity to learn about cellular heirarchies, spatial relationships, and temporal progressions within the microbiome. Weighted Gene Co-Expression Network Analysis (WGCNA) is used to find delineations in both human and mouse embryonic transcriptome dynamics during progression from oocyte to morula. [7] A similar analysis is performed using Single-cell Clustering Using Bifurcation Analysis (SCUBA), and is verified using Reverse Transcription Polymerase Chain Reaction (RT-PCR) data over the same single-cell measurements. [8] The studies conducted using network modeling approaches target single-cell sources at multiple time points, or distinct measures that could be compared using a pseudo-time mapping. Diversification of the single-cell data by incorperating multiple subjects is not considered or adressed.

Description of Motivating Example

Throughout the course of this paper, references are made to “The immune cell landscape in kidneys with lupus nephritis patients” [9]. This paper references single-cell data collected as part of a cross-sectional, case-control study of 27 Lupus Nephritis subjects. Samples of kidney tissue are taken at ten clinical sites across the United States,, where they are crygenically frozen and shiped to a central processing facility. Samples are thawed, dissociated, and sorted into single-cell suspension across 384-well plates using FlowJo 10.0.7, 11-color flow cytomery [10]. sc-RNA sequencing is performed using a modified CEL-Seq2 method [11], followed by $\sim$ 1 million paired-end reads per cell. Data can be accessed through the ImmPort repository with accession code SDY997.

Data Quality Control

The Seurat Guided Clustering Tutorial [12] is used to examine intial data and perform quality control (QC) filtering. The Seurat package allows for easy classification of low-quality observations by setting threshold values for:

1. the number of unique genes detected in each cell (nFeature), and
2. the percentage of reads that map to the mitochondrial genome (perctMT)

Item (1) is used for identifying empty or broken-cell measurements (indicated by abnormally low gene detection numbers), or duplicate/multiplicate cells measures (indicated by abnormally high gene detection numbers). Item (2) is used to identify dead and/or broken cells since dead or dying cells will retain RNAs in mitochondria, but lose cytoplasmic RNA [2].

The original dristribution of the $PerctMT$ variable across subjexts is displayed in (Figure 1) below:

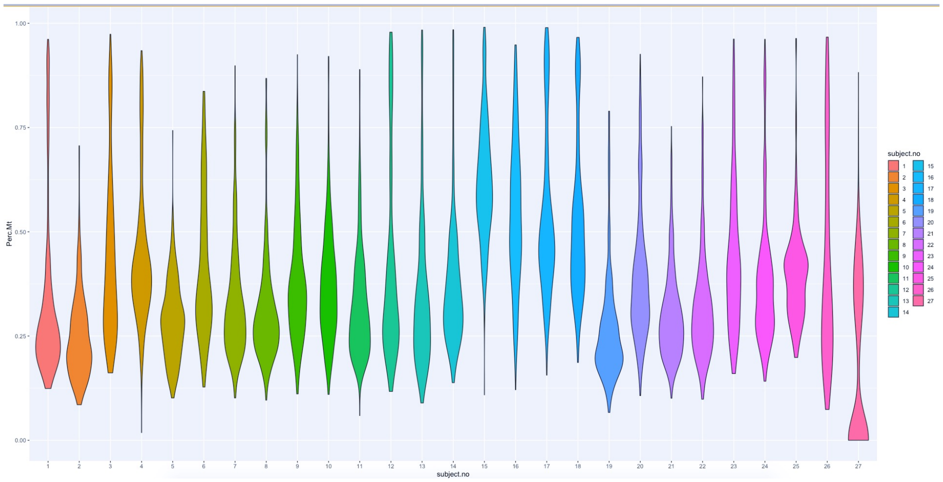

Figure 1:

The QC measures employed by (Arazi A, Rao DA, Berthier CC, et al.) and implemented using the Seurat package required:

1. $1,000 < nFeature < 5,000$
2. $perctMT \leq 25\%$

and the resulting distribution of the $PerctMT$ variable is displayed in (Figure 2):

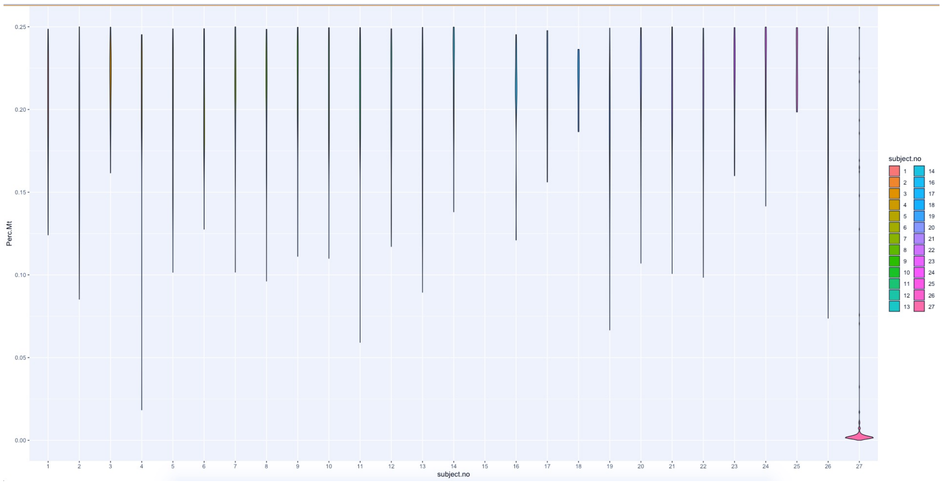

Figure 2:

a decision to increase the $perctMT$ threshold to 60% is made to preserve the inherent distribution structure across and within subjects (Figure 3).

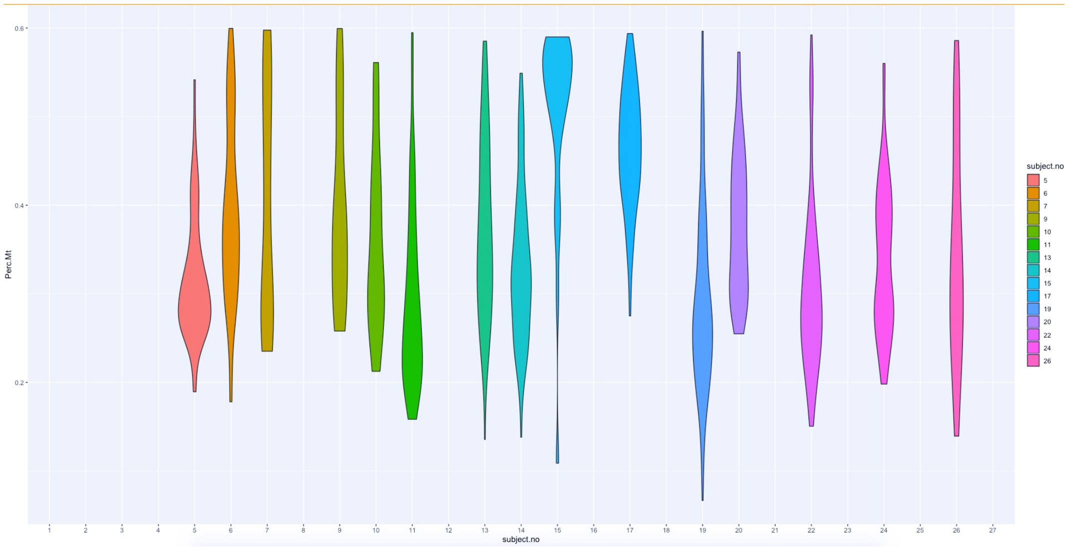

Figure 3:

Further subsetting measures are made to reduce sources of possible conflicting information, by reducing the cellular data types to B-Cells only. This will allow for a more accurate representation of the covariance parameters between-subjects since contributions of variation from inconsistency in cell-structure will be less dramatic.

The distribution of observations across subjects after the quality control thresholds are imposed is also show numerically in Table 1:

Subject Group Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Number of Observations	0	0	0	0	58	86	32	0	31	21	107	0	107	100	25	0	122	0	127	75	0	87	0	79	0	53	0

Table 1

We note that the quality control process is an active population restriction, and the data being eliminated do not constitute "missing data" under the assumption that these values poorly represented the population of interest due to innacurate measurement. As a result, subjects which lack observations can be interpreted as non-informative as opposed to missing or drop-out events. This realization will allow us to reduce the data set distribution to informative subjects, for which the observational distribution is displayed in Table 2:

Subject Group Number	5	6	7	9	10	11	13	14	15	17	19	20	22	24	26
Number of Observations	58	86	32	31	21	107	107	100	25	122	127	75	87	79	53

Table 2

Table 3 displays a 5 number summary of the observational distribution:

MIN	1st Q	Median	Mean	3rd Q	MAX
21	42.5	79	74.0	103.5	127

Table 3

Variable Selection and Summaries

In order to simplify analysis and make more significant insights into model comparisons, we choose two pairs of variables from the 38,354 genetic markers in the Lupus Data to model in a predictor-response relationship. These variables indicate higher values of correlation than arbitrary pairings, and are associated with important outcomes of interest (e.g. cancer treatment research in the case of MALAT1 [13], or observed limb malformations in the case of FBLN1 variation [14]). An attempt is also made to choose predictor-pairings of interest. The CD19 marker (paired with MALAT1) is a transmembrane protein, encoded by the CD19 gene. Since the FlowJo cytometry measurements contain CD19 protein readings, the relationship between a proteomic predictor and the outcome of interest can be modeled transitively as well as directly, which allows for more thourogh investigation of results. CD34, the predictor which we 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 data. Higher counts correspond to higher detection frequencies and (without compensating for expected expression frequency) these detection frequencies can be interpreted as a quantification of the magnitude of expression for a transcriptomic marker.

The variables that we study here are summarized in Table (4) - (8). Each describes selected variable summary statistics for subset samples specific to the subject identifiers used in Tables (1) - (3).

CD19 Summaries

Subject Number	Minimum	Maximum	Average	Median
5	0	678	36.6724	0.0
6	0	299	36.6860	7.5
7	0	10	2.1250	1.0
9	0	1052	89.4194	3.0
10	0	158	37.5714	2.0
11	0	339	28.3178	1.0
13	0	629	56.0841	18.0
14	0	251	42.2600	19.0
15	0	148	26.6000	0.0
17	0	982	112.3770	16.0
19	0	665	59.3386	5.0
20	0	287	40.1200	23.0
22	0	380	43.4483	1.0
24	0	282	55.0127	27.0
26	0	1624	268.4151	110.0

Table 4

MALAT1 Summaries

Subject Number	Minimum	Maximum	Average	Median
5	67	40812	10206.3621	9195.0
6	757	30774	11568.2791	11689.0
7	441	17916	6868	4039.5
9	311	18239	5703.9355	5983.0
10	1875	17160	6638.5714	6190.0

Subject Number	Minimum	Maximum	Average	Median
11	349	34082	9716.0280	8826.0
13	99	25572	5867.9439	4895.0
14	355	15740	6154.1500	5720.5
15	157	11923	3839.0800	3467.0
17	337	8342	2960.2541	2692.0
19	227	91961	13959.9843	10125.0
20	379	21736	7301.4133	6417.0
22	161	28429	6881.7471	5068.0
24	240	42792	6248.8228	5955.0
26	1114	32426	8463.1698	6426.0

Table 5

CD134 Summaries

Subject Number	Minimum	Maximum	Average	Median
5	0	19	3.0517	1
6	0	0	0	0
7	0	0	2	1
9	0	6	0.4516	0
10	0	5	0.6667	0
11	0	7	1.2056	1
13	0	0	0	0
14	0	1	0.4000	0
15	0	0	0	0
17	0	0	0	0

Subject Number	Minimum	Maximum	Average	Median
19	0	0	0	0
20	0	2	0.1867	0
22	0	4	0.3563	0
24	0	5	0.2911	0
26	0	0	0	0

Table 6

FBLN1 Summaries

Subject Number	Minimum	Maximum	Average	Median
5	3	41	19.3448	18
6	0	0	0	0
7	0	16	4.2500	3
9	0	8	1.8710	1
10	0	30	11.9524	10
11	0	8	1.5140	1
13	0	1	0.0093	0
14	0	5	0.5700	0
15	0	1	0.0400	0
17	0	3	0.0246	0
19	0	2	0.0157	0
20	0	9	2.5867	2
22	0	11	0.9885	0
24	0	4	0.4557	0
26	0	0	0	0

Table 7

Measurements of RNAseq data can highly specific to very precise transcriptomic targets, so while the agglomerated scope of gene expression is the same as a traditional bulk experiment, individual observations have a biologically inflated zero-component. Additionally, there are technical zero-inflation components that are associated with protocol variations.

This is evident in the case of the FBLN1 ~ CD34 pairing, where we see that expression values for several subject exhibit:

$$\underset{j}{min}(FBLN1_{ij}) = \underset{j}{min}(CD34_{ij}) = 0 = \underset{j}{max}(CD34_{ij}) = \underset{j}{max}(FBLN1_{ij})$$

where

$$i \in \{5, 6, 7, \ldots, 26\}$$
$$j \in \{1, \ldots, n_i\}$$

Which implies that:

$$(FBLN1_{ij}) = (CD34_{ij}) = 0 = (CD34_{ij}) = (FBLN1_{ij}) \quad \forall i, j$$

We expect the additional presence of zeros to be attributable to both biological and technical sources. Together, these factors contribute to heavily right-skewed variable distributions (Figure 4)

Since the MALAT1 variable had an abnormally large minimum outcome compared to the other variables, the minimum (67) outcome is subtracted from all MALAT1 values. This process would be incorperated into the model-fitting procedure automatically.

The modeling methodologies we employ motivates a log-transformation in an attempt to achieve approximate variable distribution normality, especially for the outcome variables. We

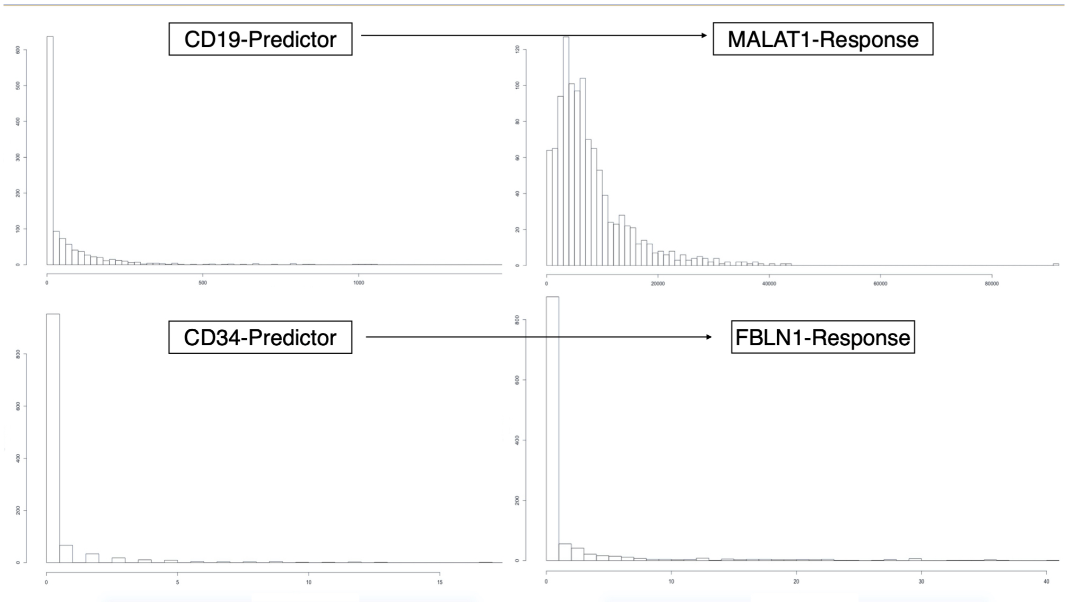

Figure 4:

perform the “log plus +1” transformation on all variables:

$$X \mapsto \log(X+1)$$

The resulting distributions are shown in Figure (5):

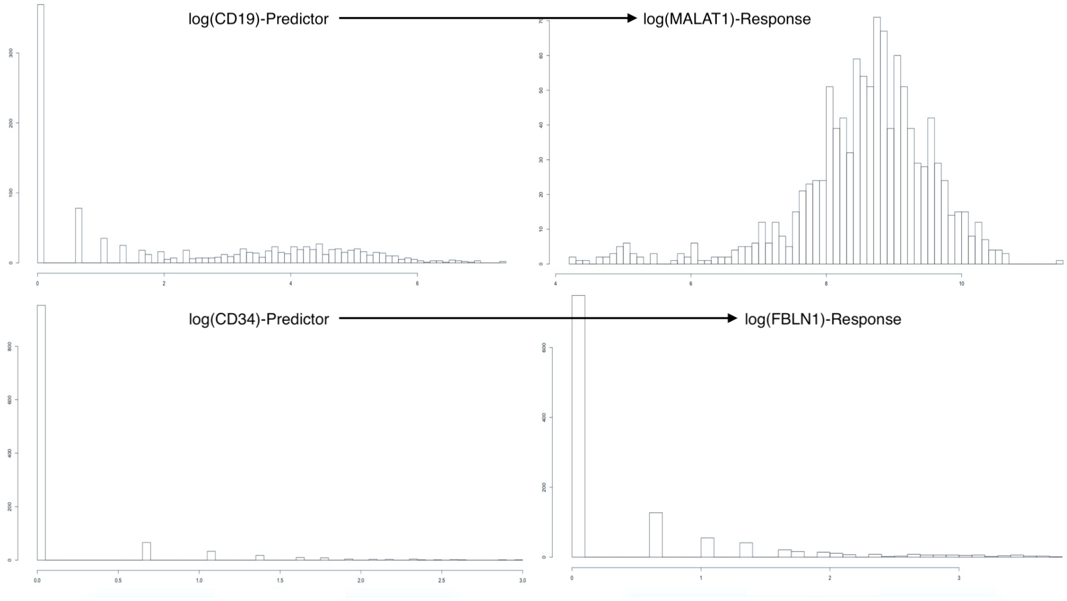

Figure 5:

We see that the log-transformed response MALAT1 has resulted in an approximately normal distribution. Conversely, the log-transformed response FBLN1 is not inherently better than the un-transformed response. We can clearly see the heavy influence of zero-inflation in these variables.

Regardless, we model each outcome under the assumption that: compensating for observational clustering will sufficiently account for non-normality of the responses. This is not generally the case, and additional transformations or modeling methodologies may be needed to improve model error distributions.

Model Descriptions

We define our outcome(s) of interest to be one of the following transformed variables as taken from (Arazi A, Rao DA, Berthier CC, et al):

$$R_h \; = \; \log\left(Y_h + 1\right) \quad \text{for} \quad h = 1, 2$$

where

$$Y_1 = \text{MALAT1-67} \quad \text{and} \quad Y_2 = \text{FBLN1}$$

We aslo define the predictor attached to R_k as:

$$P_h \; = \log\left(X_h + 1\right) \quad \text{for} \quad h = 1, 2$$

where

$$X_1 = \text{CD19} \quad \text{and} \quad Y_2 = \text{CD34}$$

Let a single response be designated as: R_{hij}. The index $i = 5, 6, \ldots, 26$ represents the subject (name of subject by number) from which the observation originated, and the index

$j = 1, \ldots, n_i$ represents the repeated observation number within subject-i. We note that $n_i \in \{21, 22, 23, \ldots, 127\}$ in the context of the Lupus Data. We present the theoretical model frameworks here as "Less Than Full Rank" (LTFR) representations. The Full-Rank model results presented create full-rank model and design matrices by droping the first level in all factors, and using this as the referrence level.

Linear Regression

We begin the model definitions by describing three linear regression models, with parameters estimated using Least Squares optimization. It should be noted that these methods make the assumption that observations are independent, and should therefore be used for comparison to modeling methods to come. However, the linear regression models we present here can account for some observational clustering with the use of subject specific intercept and slope terms.

Ultimately, though, all the methods defined in this section assume an identical error structure across all observations of the form:

$$\epsilon_{hij} \sim N\left(0, \sigma_\epsilon^2 * I_{1110}\right)$$

where we are assuming that σ^2 is a common variance parameter for all subjects and I_{1110} is the 1110 X 1110 identity matrix.

Simple Linear Regression (Model 0)

Using the notation we defined above, we write the first model as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + \epsilon_{hij}$$

which is equivalent to:

$$\log(Y_{hij}) = \beta_0 + \beta_1 \log(X_{hij}) + \epsilon_{hij}$$

We note that this model does not account for any observational clustering.

Fixed-Effect Subject-Intercept (Model 1)

Adding a subject-specific intercept term, allows us to account for within-subject correlation by uniformly shifting the fitted values specific to a subject. This model may be written as:

$$R_{hij} = \beta_0 + \beta_{1i}(subject_i) + \beta_2 P_{hij} + \epsilon_{hij}$$

where we define the term:

$$\beta_{1i}\left(subject_i\right) = \begin{cases} \beta_{1i} & \text{if} \quad subject_i = i \\ 0 & \text{if} \quad subject_i \neq i \end{cases}$$

Fixed-Effect Subject-Slope (Model 2)

We may further account for observational clustering by adding a term which will ensure that individual subjects' relationships with the covariate of interest is accounted for. This will help to reduce within-subject variation across the predictor space, and will be more noticable for stronger, subject-specific interactions with covariates.

This model may be written as:

$$R_{hij} = \beta_0 + \beta_{1i}\left(subject_i\right) + \left[\beta_{2i}\left(subject_i\right) * P_{hij}\right] + \beta_3 P_{hij} + \epsilon_{hij}$$

where we are using the same definitions of $(subject_i)$, R_{hij}, and P_{hij} as in Models 0 and 1.

Linear Mixed Effects Models

The next category of modeling approached we describe is Linear Mixed Effect Models. Specifically, we describe two distinct Linear Mixed Effect Models that account for subject-clustering differently than the previously discussed Linear Regression models. Linear Mixed Efffects Models do not neccessarily assume observational independence. Correlation structures such as AR(1), independence, spatial power, or unstructured can be used to estimate parameters determining covariance amongst repeated measures within a subject and between observations across subjects. Additionally, if we can rationally assume that the responses shown in Figure 3 have a multivariate distribution, the model parameters can be easily estimated using Maximum Likelihood Estimation techniques [15].

Linear Mixed Effects Model with Random Intercept (Model 3)

Model 1 accounts for subject-clustering by assuming that observations within a subject are uniformly influenced by the nested nature (observations within subjects) of the sampling method. However, this assumption may not always be reasonable, as we could imagine that responses within each subject exhibit random variation that is also related to nested sampling methods.

A Linear Mixed Effects Model that includes a Random Intercept accounts for observational covariation due to subject-clustering by assuming that observations within a subject are a consequence of the nested nature of the sampling method, and therefore a consequence of an additive (covariate-independent), subject-specific, effect; AND due to subject-specific random variation in response measurement associated with measurement instatbility for THAT subject.

This model may be written as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + b_{0i}\left(subject_i\right) + \epsilon_{hij}$$

where

$$b_{0i} \sim N\left(0, \sigma_b^2\right)$$

$$\epsilon_{hij} \sim N\left(0, \sigma_\epsilon^2 I_{n_i}\right)$$

and we assume that b_{0i} and ϵ_{hij} are independent.

We note that both random-components can be assumed to have a mean of zero as non-zero components are inherently deterministic and can be integrated into intercept terms.

Linear Mixed Effect Model with Random Slope (Model 4)

Model 2 implements a Fixed Effect slope in an attempt to reconcile the effects of observational clustering that was inadequately accounted for by the subject-specific Fixed Effect Intercept in Model 1. However, in light of the information surrounding the development of Model 3, it is incumbent for us to develop an analogous correction for Model 2. Such a correction will allow us to account for observational correlation due to subeject-clustering as sourced from:

- additive, effects due to subject-clustered nested sampling methods
- subject-specific random variation associated with measurement instability
- covariate-dependent, subject-specific effects
- covariate-dependent, subject-specific random variation associated with measurement instability

We write this model as:

$$R_{hij} = \beta_0 + \beta_1 P_{hij} + b_{0i}\left(subject_i\right) + \left[b_{1i}\left(subject_i\right) P_{hij}\right] + \epsilon_{hij}$$

where

$$\mathbf{b} = \begin{bmatrix} b_{0i} \\ b_{1i} \end{bmatrix} \sim N\left(\mathbf{0}, \mathbf{G}\right)$$

$$G = \begin{bmatrix} \sigma_b^2 & 0 \\ 0 & \sigma_b^2 \end{bmatrix}$$

$$\epsilon_{hij} \sim N\left(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}_{n_i}\right)$$

Generalized Estimating Equations (Model 5)

Our final method for modeling scRNA-seq expression profiles is Generalized Estimating Equations (GEE). Dissimilar to each of the methods previously described, GEE regression esitimates are obtained using methodologies that allow for non-continuous responses. GEE extrapolates on the techniques used for estimating Generalized Linear Models by incorperating the effects of observational correlation and clustering.

GEE estimates are computed by solving the estimating equation(s):

$$0 = U(\beta) = \sum_{i=1}^{15} \left\{ \mathbf{D}_{hi}^T \mathbf{V}_{hi}^{-1} \left(\mathbf{y}_{hi} - \mu_{hi} \right) \right\} \tag{1}$$

where:

$$\mu_{hi} = \mu_{hi}(\beta) = E\left[\mathbf{Y}_{hi}\right] = \eta_{hi}$$

represents the relationship between the expected value of the response μ_i (not necessarily assumed to be related to a distribution) and the linear predictor η_i,

$$\mathbf{D}_{hi} = \begin{bmatrix} \frac{\partial\mu_{hi1}}{\beta_1} & \frac{\partial\mu_{hi1}}{\beta_2} & \cdots & \frac{\partial\mu_{hi1}}{\beta_p} \\ \frac{\partial\mu_{hi2}}{\beta_1} & \frac{\partial\mu_{hi2}}{\beta_2} & \cdots & \frac{\partial\mu_{hi2}}{\beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial\mu_{hin_i}}{\beta_1} & \frac{\partial\mu_{hin_i}}{\beta_2} & \cdots & \frac{\partial\mu_{hin_i}}{\beta_p} \end{bmatrix}$$

is the first derivative matrix, and

$$\mathbf{V}_{hi} = \mathbf{A}_{hi}^{\frac{1}{2}} Corr(\mathbf{Y}_{\mathbf{hi}}) \mathbf{A}_{hi}^{\frac{1}{2}}$$

$$\mathbf{A}_{hi} = \underset{n_i}{diag} \left\{ \phi_j \left(t_{ij} \right) \nu \left(\mu_{hij} \right) \right\}$$

We note that $\phi_j \left(t_{ij} \right)$ and $\nu \left(\mu_{hij} \right)$ are hyperparameters defined so that we may know the variance as a function of the mean and a scale parameter, i.e:

$$Var \left(Y_{hij} \right) = \phi_j \left(t_{ij} \right) \nu \left(\mu_{hij} \right)$$

The GEE algorithm is iterative and used the following steps to converge at an estimate:

1. Generalized Linear Modeling methods employing Maximum Likelihood Estimation are used to obtain intial estimates for β
2. Estimates for β used to compute hyper-parameters
3. New estimates for hyper-parameters and working covariance matrix $(\mathbf{V}_{hi})$ used to obtain new estimates for β by solving (1)
4. Repeat Steps 2 & 3 until algorithm converges

The GEE algorithm has a quality which makes it very appealing for many applications with observational clustering. Specifically, the algorithm is robust to misspecification of the observational correlation structure. That is, the estimates $\hat{\beta}_{GEE}$ are consistent with β irrespective of the estimates for within-subject correlation.

The stability of the GEE algorithm is in-part due to the effects that it estimates. Whereas each of the previous methods (Model 0 withstanding) had subject-specific interpretations, the GEE algorithm provides marginal parameter estimates. These values do not represent any specific subject, but rather the population-average.

According to (Fitzmaurice GM, Laird NM, Ware JH (2012)) [15] we also need to ensure that any responses modeled in the GEE process are stationary, i.e:

$$E\left[Y_{hij}|\mathbf{X}_{hi}\right] = E\left[Y_{hij}|X_{hi1}, \ldots, X_{hin_i}\right] = E\left[Y_{hij}|X_{hij}\right]$$

The scRNA-seq data has been assumed to be independent within-subject, therefore we have:

$$E\left[Y_{hij}|X_{hij}\right] = E\left[Y_{hij}|X_{hij'}\right]$$

$$\forall j \in \{1, \ldots, n_i\} \quad j \neq j'$$

Since the use of the scRNA-seq data would not violate the GEE assumptions, we proceed with the description of the model that we will fit.

The three-part specification includes:

1. The link function and linear predictor
2. Variance function
3. A working covariance matrix

The link function and linear predictor are chosen so that the resulting model estimates will be comparable to preceeding estimates for intercept and slope. Therefore, we will use the identity link function in conjunction with the linear predictor:

$$g(x) = x$$

$$g(\mu_{hij}) = \eta_{hij} = \beta_0 + \beta_1 P_{hij}$$

which implies we will be assuming the general modeling structure:

$$E\left[Y_{hij}\right] = \mu_{hij} = \eta_{hij} = \beta_0 + \beta_1 P_{hij}$$

In the abscence of further information, we will assume a variance function of the form:

$$Var\left(Y_{hij}\right) = \phi$$

and we will be using a working covariance matrix structure for repeated measures that corresponds to the assumption of independence of observations within a subject.

$$\left[Corr\left(Y_{hij}, Y_{hik}\right)\right]_{jk} = \begin{cases} 1 \quad \text{if} \quad j = k \\ \\ 0 \quad \text{if} \quad j \neq k \end{cases}$$

$$for \quad j, k \in \{1, \ldots, n_i\}$$

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 necessarry and referrence 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. 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.

4. Amir E-aD, Davis KL, Tadmor MD, et al. (2013) ViSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature biotechnology* 31: 545.

5. Wang Z, Gerstein M, Snyder M (2009) RNA-seq: A revolutionary tool for transcriptomics. *Nature reviews genetics* 10: 57.

6. Kharchenko PV, Silberstein L, Scadden DT (2014) Bayesian approach to single-cell differential expression analysis. *Nature methods* 11: 740.

7. Xue Z, Huang K, Cai C, et al. (2013) Genetic programs in human and mouse early embryos revealed by single-cell rna sequencing. *Nature* 500: 593.

8. Marco E, Karp RL, Guo G, et al. (2014) Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences* 111: E5643–E5650.

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

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

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

12. Satija R, others (2018) Seurat: Guided clustering tutorial. *Satija Lab http://satijalab org/seurat/pbmc3k_tutorial html.*

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

14. 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.

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