Differential Expression Analysis Techniques for Single-Cell RNA-seq Experiments

for the Computational Biology Doctoral Seminar (CMPBIO 293),
organized by N. Yosef & T. Ashuach, Spring 2018, UC Berkeley

Kevin Benac and Nima Hejazi

Group in Biostatistics,
University of California, Berkeley

11 April 2018

Outline

- Introduction
  - Data
  - Objective
- Methodology
  - ZINB-WaVE
  - DropLasso
- Conclusions
  - Comparison

The Data: Single-Cell RNA-seq

- scRNA-seq fast growing approach to measure the genome-wide transcriptome of many individual cells in parallel (Kolodziejczyk et al., 2015).
- Major advance compared to standard bulk RNA sequencing to investigate complex heterogeneous tissues,
- Access to cell-to-cell variability: better accuracy.

The Data: Single-Cell RNA-seq

- However, analysis of single-cell RNA-seq data is challenging.
- In one cell, only a tiny amount of RNA is present and large fraction of polyadenylated RNA can be stochastically lost during sample preparation steps (cell lysis, reverse transcription or amplification).
  $\Longrightarrow$ Many genes fail to be detected although they are expressed!
- In practice, not uncommon to end up with a matrix of read counts where about 80% of the coefficients are zeros.
- This zeros are called *dropouts*.

The Data: Single-Cell RNA-seq

	Cell1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	Cell 7
Xkr4	0	0	0	14	0	0	0
Syt11	1	9	2	2	0	0	0
Cpe	0	0	16	0	0	0	0
Rp1	0	0	0	0	0	0	0
Gm73	0	0	0	0	0	0	0
Gm79	0	0	0	0	0	0	0
Mpl15	8	8	6	1	0	0	0
Gm61	0	0	0	0	0	3	0
Lypla1	1	23	266	1	0	1	0
Tcea1	63	101	18	29	2	34	0

The Data: Single-Cell RNA-seq

- Raises modelling and computational issues.
- Need to detect a signal when most of the values are zeros only because they are missing.
- Traditional methods used for bulk RNA-seq data might not be sensible anymore.

Outline

Introduction

- Data
- Objective

Methodology

- ZINB-WaVE
- DropLasso

Conclusions

- Comparison

The Objective: Differential Expression

- Why “differential”? The goal is to find a subset of relevant biomarkers with respect to a particular condition of interest (e.g., disease, tissue of origin).

- Many experimental settings seek to isolate a subset of biomarkers from the full (larger) assayed set in order to identify biological patterns and better inform future biological experiments.

- Since experimental costs are high and modern biotechnologies allow numerous biological targets (e.g., genes) to be assayed, the result is a very high-dimensional statistical problem.

The Objective: Differential Expression

- Regularized Linear Models:

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(w, x_i, y_i) + \lambda \Omega(w) \right\}$$

- Lasso for continuous outcomes (squared-error loss):

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} w_j x_{i,j} \right)^2 + \lambda \sum_{j=1}^{d} |w_j| \right\}$$

Outline

Introduction

- Data
- Objective

Methodology

- ZINB-WaVE
- DropLasso

Conclusions

- Comparison

ZINB-WaVE

- Method that leads to low-dimensional representations of the data the same way PCA or tSNE does.

ZINB-WaVE

- Method that leads to low-dimensional representations of the data the same way PCA or tSNE does.
- However accounts for zero inflation (dropouts), over-dispersion, and the count nature of the data.
- No need for normalization.

ZINB-WaVE

Mathematical set-up:

- n samples (single-cells),
- J genes,
- Y_{ij} read counts for gene j in cell i, $1 \leq \ldots \leq n, \quad 1 \leq j \leq J.$,
- π_{ij}: probability of dropout,
- μ_{ij}: mean expression level.

ZINB-WaVE

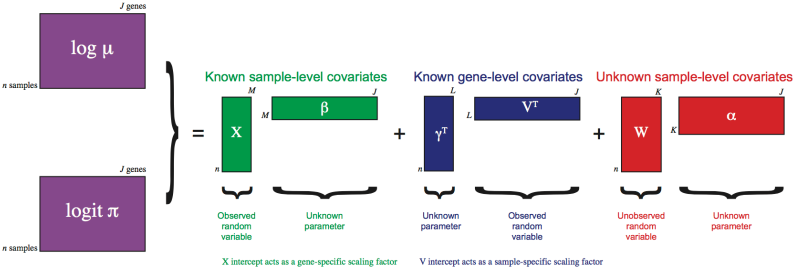

J genes
log μ
n samples
J genes
logit π
n samples
=
Known sample-level covariates
M
J
M
X
β
n
+
Known gene-level covariates
L
J
L
γ^T
V^T
n
+
Unknown sample-level covariates
K
J
K
W
α
n
Observed random variable
Unknown parameter
Unknown parameter
Observed random variable
Unobserved random variable
Unknown parameter
X intercept acts as a gene-specific scaling factor
V intercept acts as a sample-specific scaling factor

Figure 1: The ZINB-WaVE model

ZINB-WaVE

- ZINB-WaVE mainly used for normalization and dimensionality reduction but can also be used for DE analysis.
- Compute weights from the estimated π using Bayes formula.
- If the observed counts are positive, $w = 1$, otherwise, $0 < w < 1$.
- The higher π, the lower w

ZINB-WaVE

- Once we have the weights, fit a weighted negative binomial generalized linear model using the ZINB-WaVE weights.
- End-up with a matrix of fitted values.
- Not sparse anymore, look more like bulk RNA-seq data. $\Longrightarrow$ We can use classical tools for differential expression analysis (e.g. edgeR, DESeq2, limma-voom in R/Bioconductor).

Outline

Introduction
- Data
- Objective

Methodology
- ZINB-WaVE
- DropLasso

Conclusions
- Comparison

DropLasso

- Consider the following data structure:
  - $x_i \in \mathbb{R}^d$ — design matrix of scRNA-seq counts
  - $y_i \in \mathbb{R}$ — cell-level outcome of interest (e.g., tissue of origin)
  - $\delta_i \in \{0, 1\}^d$ s.t. $\delta_i \sim Bern(p)^d$ — random dropout mask
  - $\delta \odot x \in \mathbb{R}^d$ — corrupted pattern for scRNA-seq dropout
  - $\mathrm{P}(\delta_i = 1) = \mathrm{p}$ — probability of *not* being censored by dropout
- The DropLasso procedure seeks to identify differentially expressed genes based on cell-level differences while accounting for the dropout noise that masks scRNA data.

DropLasso

- Introducing dropout ($\delta_i \sim Bern(p)^d$):

$$\min_{w \in \mathbb{R}^d} \left\{ \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\delta_i} \mathcal{L}\left(w, \delta_i \odot \frac{x_i}{p}, y_i \right) + \lambda \|w\|_1 \right\}$$

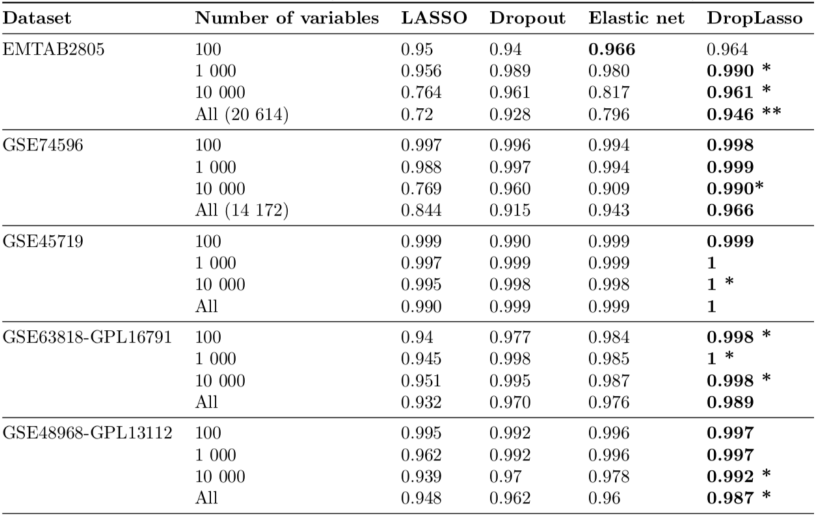
- Independence from p in expectation:

$$\begin{aligned} \mathbb{E}_{\delta_i} \sum_{j=1}^{d} w_j \left(\delta_i \odot \frac{x_i}{p} \right)_j &= \sum_{j=1}^{d} \mathbb{E}_{\delta_i} w_j \delta_{i,j} \frac{x_{i,j}}{p} \\ &= \sum_{j=1}^{d} w_j x_{i,j} \end{aligned}$$

DropLasso

- Introducing the dropout term δ amounts to censoring the observed data and adjusting (i.e., $\frac{x_p}{p}$) such that the effects of dropout noise are removed.
- This places a *statistical model* on the dropout noise — i.e., $\delta_i \sim Bern(p)^d$
  - Dropout noise is independent across samples and genes. (Fine starting point but probably untrue scientifically.)
  - Modeling dropout noise in a more flexible manner could likely improve DropLasso performance and is identified as an item of future work.
- Merely introducing the simple dropout correction significantly improves performance under standard modeling metrics (e.g., AUC).

DropLasso



Dataset	Number of variables	LASSO	Dropout	Elastic net	DropLasso
EMTAB2805	100	0.95	0.94	**0.966**	0.964
	1 000	0.956	0.989	0.980	**0.990** *
	10 000	0.764	0.961	0.817	**0.961** *
	All (20 614)	0.72	0.928	0.796	**0.946** **
GSE74596	100	0.997	0.996	0.994	**0.998**
	1 000	0.988	0.997	0.994	**0.999**
	10 000	0.769	0.960	0.909	**0.990***
	All (14 172)	0.844	0.915	0.943	**0.966**
GSE45719	100	0.999	0.990	0.999	**0.999**
	1 000	0.997	0.999	0.999	**1**
	10 000	0.995	0.998	0.998	**1** *
	All	0.990	0.999	0.999	**1**
GSE63818-GPL16791	100	0.94	0.977	0.984	**0.998** *
	1 000	0.945	0.998	0.985	**1** *
	10 000	0.951	0.995	0.987	**0.998** *
	All	0.932	0.970	0.976	**0.989**
GSE48968-GPL13112	100	0.995	0.992	0.996	**0.997**
	1 000	0.962	0.992	0.996	**0.997**
	10 000	0.939	0.97	0.978	**0.992** *
	All	0.948	0.962	0.96	**0.987** *

Figure 2: Excerpt from table 3 of "DropLasso: A robust variant of Lasso for single cell RNA-seq data" Khalfaoui & Vert (2018)

Outline

Introduction

- Data
- Objective

Methodology

- ZINB-WaVE
- DropLasso

Conclusions

- Comparison

ZINB-WaVE v. DropLasso

- ZINB-WaVE is designed to address issues in the statistical analysis pipeline that come before differential expression analysis:
  - Normalization
  - Dimensionality Reduction
- Since ZINB-WaVE attempts to make scRNA-seq data resemble bulk RNA-seq data, the weights can be used with standard differential expression tools.

ZINB-WaVE v. DropLasso

- DropLasso seeks to cast the scRNA-seq DE problem as a standard Lasso problem, accounting for dropout noise using the regularization introduced in the neural networks literature.

- Since DropLasso is a very new method, there have been no in-depth comparisons of the two techniques as of yet.

References I

Beyrem Khalfaoui and Jean-Philippe Vert. DropLasso: A robust variant of Lasso for single-cell RNA-seq data. *arXiv preprint arXiv:1802.09381*, 2018.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. *bioRxiv*, 2017.