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

Kevin Benac and Nima Hejazi

Group in Biostatistics,
University of California, Berkeley

11 April 2018

Outline

The Data: Single-Cell RNA-seq

- ...

The Data: Single-Cell RNA-seq

- ...

The Data: Single-Cell RNA-seq

- ...

Outline

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

ZINB-WaVE I

- ...

- ...

ZINB-WaVE II

- ...

- ...

ZINB-WaVE III

- ...

- ...

ZINB-WaVE IV

- ...

- ...

Outline

DropLasso I

- 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

  - $P(\delta_i = 1) = p$ — 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 II

- 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:

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

$$= \sum_{j=1}^{d} w_j x_{i,j}$$

DropLasso III

- 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 IV

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

Outline

ZINB-WaVE v. DropLasso I

- ...
- ...

ZINB-WaVE v. DropLasso II

- ...

- ...

ZINB-WaVE v. DropLasso III

- ...
- ...

Outline

Review

- ...

- ...

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.