Advantages of Modeling Zero-Inflation in scRNA-Seq Data

Davide Risso*

Abstract
Single-cell RNA sequencing is a novel technique that allows researchers to measure gene expression at the resolution of single cells. Compared to "bulk" measurements, single-cell data show an over-abundance of zero counts, and several statistical models have been proposed to account for this zero inflation. Here, we show that in the context of dimensionality reduction, a negative binomial factor analysis model leads to similar results than its zero-inflated counterpart, with substantial computational savings. However, explicitly testing for the difference in the proportion of extra zeros may help identify interesting genes.

Key Words: single cell, RNA-seq, zero inflation, count data, dimensionality reduction, factor analysis

1. Introduction

The advent of single-cell RNA sequencing (scRNA-seq) has revolutionized the field of transcriptomics, providing for the first time the tools to disentangle the heterogeneity of complex tissues (Wagner, Regev, and Yosef 2016).

The negative binomial has emerged as the distribution of choice for the modeling of RNA-seq gene-level read counts. Indeed, several authors have shown how technical replicates of the same biological unit vary according to Poisson noise (Marioni et al. 2008; Bullard et al. 2010), while the added biological variability causes over-dispersion (Robinson, McCarthy, and Smyth 2010; Anders and Huber 2010). The negative binomial distribution, which can be seen as a Gamma-Poisson mixture, leading to a quadratic mean-variance relation, very well fits the observed data in a wide variety of contexts.

From an interpretation point of view, this is best seen by looking at the square of the coefficient of variation:

$$CV^2 = \frac{1}{\mu} + \phi,$$

where μ and ϕ are the mean and dispersion parameters of the negative binomial, respectively. The first term, which tends to zero as the number of reads increases, represents the technical variability due to the sequencing, while the second term, independent from the mean, represents the biological variability, a property of the system under study (McCarthy, Chen, and Smyth 2012).

However, compared to the so-called "bulk" measurements, single-cell data exhibit higher variance, expression outliers, and an over-abundance of zero counts (Marinov et al. 2014; Cole et al. 2019). A variety of different approaches have been published to tackle the problem of over-abundance of zeros. Kharchenko, Silberstein, and Scadden (2014) proposed a Bayesian mixture of a low-mean Poisson and a negative binomial for the differential expression of single-cell data; Finak et al. (2015) used a Hurdle Gaussian model for the same application; Pierson and Yau (2015) and Risso et al. (2018) tackled the problem of dimensionality reduction, using factor analysis models.

*Department of Statistical Sciences, University of Padova, via C. Battisti 241, 35121 Padova, Italy; Division of Biostatistics and Epidemiology, Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY, USA.

In an effort to reduce the number of expression outliers, largely due to the so-called "amplification bias", a unique molecular identifier (UMI) can be added to each individual molecule prior to the amplification step of the library preparation (Islam et al. 2014). In addition to reducing the technical variability, it was recently shown that the inclusion of UMIs in the protocol can reduce the effect of zero counts on the goodness of fit of the negative binomial distribution (Townes et al. 2019; Svensson 2019).

Here, we discuss the advantages and disadvantages of explicitly modeling the extra zero counts observed in real single-cell data, with particular focus on dimensionality reduction and differential expression.

2. Factor Analysis Models for Count Data

We focus our discussion on the factor analysis model of Risso et al. (2018), which employs a zero-inflated negative binomial (ZINB) model to account for the count nature of the data while explicitly modeling the extra zeros observed in real datasets.

Given $i = 1, \dots, n$ samples and $j = 1, \dots, p$ features, let us denote with Y the $n \times p$ matrix, whose (i, j) element is the realization of a random variable representing the observed count of feature j in sample i. We assume that Y follows a ZINB distribution, i.e.

$$p_{\text{ZINB}}(y; \mu, \theta, \pi) = \pi \delta_0(y) + (1 - \pi) p_{\text{NB}}(y; \mu, \theta), \quad \forall y \in \mathbb{N}, \tag{1}$$

where $\delta_0(\cdot)$ is the Dirac function, $\pi \in [0, 1]$ is the mixing parameter, and $p_{\text{NB}}(y; \mu, \theta)$ is the negative binomial distribution with mean parameter $\mu \geq 0$ and dispersion parameter $\theta > 0$:

$$f_{\text{NB}}(y; \mu, \theta) = \frac{\Gamma(y + \theta)}{\Gamma(y + 1)\Gamma(\theta)} \left(\frac{\theta}{\theta + \mu}\right)^{\theta} \left(\frac{\mu}{\mu + \theta}\right)^{y}, \quad \forall y \in \mathbb{N}. \tag{2}$$

As in Risso et al. (2018), we specify the following regression models for the parameters:

$$\log(\mu_{ij}) = \left(X\beta_\mu + (V\gamma_\mu)^\top + W\alpha_\mu\right)_{ij}, \tag{3}$$

$$\text{logit}(\pi_{ij}) = \left(X\beta_\pi + (V\gamma_\pi)^\top + W\alpha_\pi\right)_{ij}, \tag{4}$$

where where X is a $n \times M$ matrix containing M sample-level observed covariates, V is a $p \times Q$ matrix containing Q feature-level observed covariates, and W is a $n \times K$ matrix of unobserved factors, which can be interpreted as explaining the original data in fewer dimensions (see Risso et al. (2018) for details).

The ZINB distribution is particularly appealing since it reduces to the negative binomial if the data are not truly zero inflated and it is able to discriminate different types of zeros. In fact, some zeros may be "biological", i.e., genes that are simply not expressed in the cell under study, or "technical", i.e., genes that are expressed but not captured by the technology. Note that this is a fundamental difference between zero-inflated models and Hurdle models, which assume that all the zeros are coming from a separate component. In zero-inflated models, instead, the structural zeros (in this case called biological zeros) should be captured by the point-mass at zero and the sampling zeros (in this case called technical zeros) are modeled by the negative binomial distribution.

If zero inflation is not evident in the data at hand, it might be beneficial to consider a more parsimonious model, in which Y follows a negative binomial distribution with parameters $\mu_{ij} \geq 0$ and $\theta_j > 0$ and

$$\log(\mu_{ij}) = \left(X\beta + (V\gamma)^\top + W\alpha\right)_{ij}, \tag{5}$$

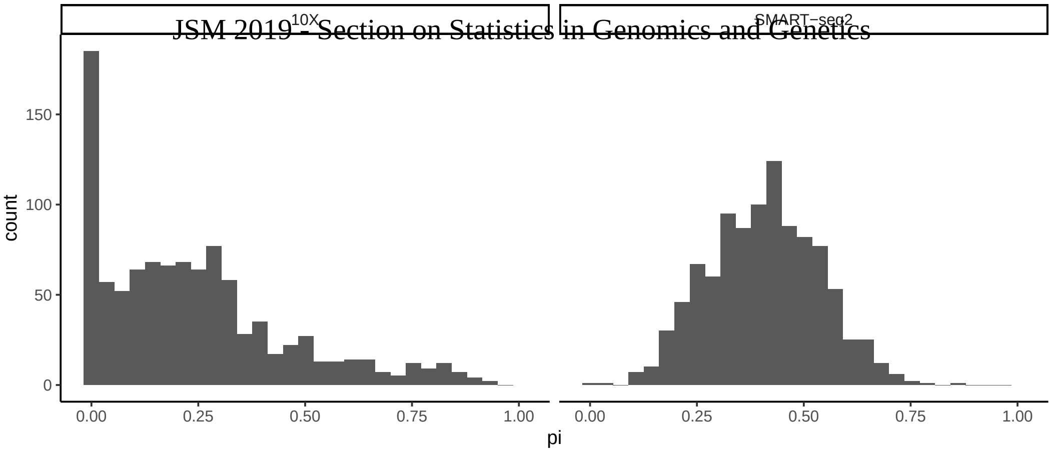

Figure 1: Per-gene distribution of the estimates of π. Left panel: 10X data. Right panel: Fluidigm data.

where X, V, and W are the same as in equation (3). The number of parameters of the model in equation (5) is $J(M+K+1)+n(Q+K)$, $J(M+K)+nQ$ fewer than the model in (3).

2.1 Parameter estimation

To infer the parameters, we used a numerical algorithm that maximizes the penalized likelihood, as done in Risso et al. (2018).

Briefly, we solve

$$\max_{\beta,\gamma,W,\alpha,\zeta} \left\{ \ell(\beta,\gamma,W,\alpha,\theta) - \text{Pen}(\beta,\gamma,W,\alpha,\theta) \right\} ,$$

where $\ell(\cdot)$ is the log-likelihood function and $\text{Pen}(\cdot)$ is a regularization term to reduce overfitting and improve the numerical stability of the optimization problem in the setting of many parameters. See Risso et al. (2018) for details.

The likelihood only depends on W and α through their product $R = W\alpha$ and the penalty ensures that at the optimum W and α have orthogonal columns, which is useful for visualization or interpretation of latent factors (Lemma 1 of Risso et al. (2018)).

3. Datasets Used

For the analyses in the next section, we used two publicly available scRNA-seq datasets. The first, referred to as the 10X dataset, was created by concatenating two sets of peripheral blood monocytes (PBMCs) coming from two healthy individuals, available in the *TENxPBMCData* Bioconductor package (Hansen, Risso, and Hicks 2019). In particular, the datasets *frozen_pbmc_donor_a* and *frozen_pbmc_donor_b* were used; by using two donors we expected to observe a "batch effect" and we used this dataset to test the ability of the models of removing such effect.

The second dataset, referred to as the Fluidigm dataset, was retrieved from the *scRNAseq* Bioconductor package (Risso, Cole, and Lun 2019) and contains a set of maturing neurons and neuronal progenitor cells described in Pollen et al. (2014). The dataset includes 65 cells sequenced at two different depths: the "low-coverage" data have an average of about 90,000 reads per cell, while the "high-coverage" data have an average of about 3 million reads per cell.

Of note, the first dataset employs UMIs to reduce the amplification bias, while the second does not. Hence, we expected a more dramatic zero inflation in the Fluidigm data than in the 10X data.

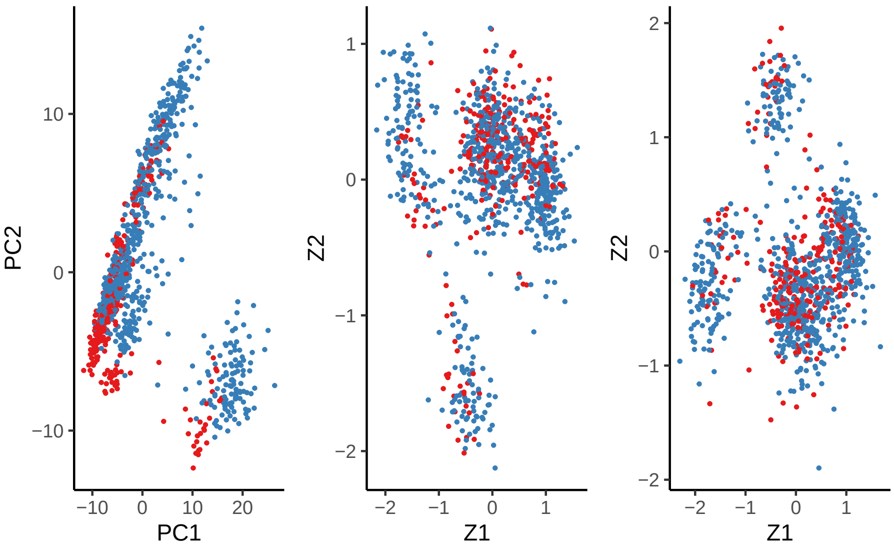

Figure 2: Low-dimensional representation of the 10X data ($K = 2$). Left panel: PCA; center panel: NB; right panel: ZINB. Red points are cells coming from *frozen_pbmc_donor_a*, while blue points are cells from *frozen_pbmc_donor_b*.

4. Results

4.1 Zero inflation in single-cell data

Recently, Townes et al. (2019) and Svensson (2019) independently observed that in some single-cell data, particularly those that employ UMIs, the abundance of zeros (which can be as large as 80% of the data) can be fully explained by the negative binomial distribution, with no need to treat zeros differently than any other count. Hence, UMIs, while not affecting the proportion of zero counts, indirectly impact the fit of the negative binomial. In fact, by construction, UMI counts are lower than read counts and while the number of zeros stays the same in the two cases, the negative binomial distribution with a lower mean is much more likely to explain the high proportion of observed zeros.

To confirm this observation, we fitted the ZINB model of equations (3) and (4) to both a UMI and a non-UMI dataset. The per-gene distribution of the resulting estimated π is shown in Figure 1: the difference between the two datasets is evident, with the UMI data showing a lack of zero inflation ($\hat{\pi} = 0$) for a large fraction of genes, and the non-UMI data showing zero inflation for virtually all genes.

4.2 Effects on Dimensionality Reduction

We then looked at the effect of zero inflation on dimensionality reduction. We compared the performance of the ZINB and negative binomial (NB) factor models, as well as of a naive procedure that applies principal component analysis (PCA) on the logarithm of normalized counts.

Compared to PCA, both the ZINB and NB models were able to remove the batch effects (included as a covariate in X) and generally led to more distinct and well-behaved clusters of cells (Figure 2). To evaluate the dependence of the inferred low-dimensional signal on technical variation, we computed the absolute correlation between each dimension (e.g.,

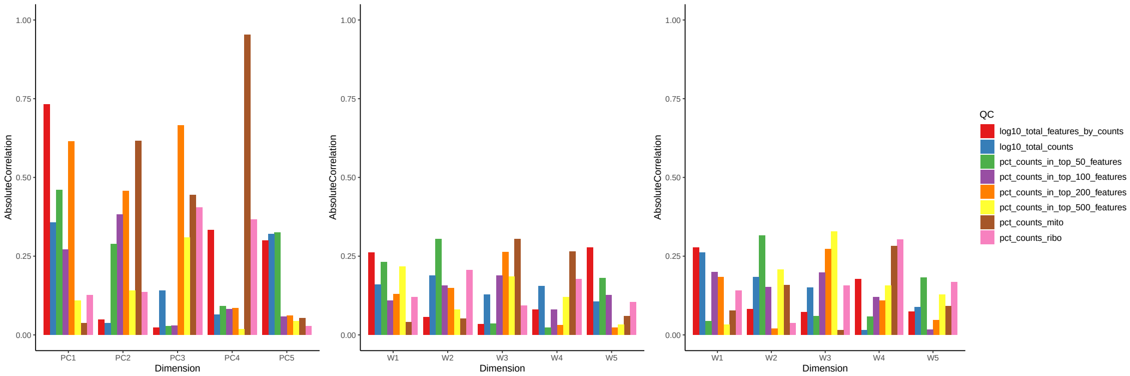

Figure 3: Absolute correlation between QC measures and the first 5 factors inferred by the models. Left panel: PCA; center panel: NB; right panel: ZINB.

principal component) and a set of quality control (QC) measures, computed with the *scater* Bioconductor package (McCarthy et al. 2017). While PCA showed high correlation with several of these metrics, both the NB and ZINB model did not, suggesting that the inferred signal represents biological, rather than technical, variation (Figure 3).

4.3 Effects on Differential Expression

The models described in equations (3 - 5) can be used to detect differentially expressed genes. To do so, it is sufficient to include an indicator variable (or a set of dummy variables) representing the groups to compare as a column of X and to test the null hypothesis that the corresponding element of β_μ and/or β_π is equal to zero.

Clearly, the NB model is limited to testing the coefficient of the regression on the negative binomial mean. The ZINB model, on the other hand, can test both the effect of the variable of interest on the mean count and on the proportion of zeros. The latter can be useful to identify those genes that are expressed only in a subset of cells per condition.

For illustration, we employ a Bayesian approach to identify differentially expressed genes. In particular, we fit a simpler version of the models without the $W\alpha$ terms and we assume independent normal priors for the parameters, i.e.

$$\begin{aligned}
\beta_\mu^j &\sim N(0, \tau_\beta);\\
\beta_\pi^j &\sim N(0, \tau_\beta);\\
\gamma_\mu^i &\sim N(0, \tau_\gamma);\\
\gamma_\pi^i &\sim N(0, \tau_\gamma);\\
\log \theta^j &\sim N(0, \tau_\theta).
\end{aligned}$$

The model was implemented in RStan (Stan Development Team 2019) to obtain the posterior distribution of the parameters. In order to identify differential expression on the proportion of zero counts, the null hypothesis $H_0 : \beta_\pi^j = 0$ was tested for each $j = 1, \dots, p$. The Bayesian False Discovery Rate approach of Van De Wiel et al. (2013) was used to identify significant genes.

To check whether the model is able to identify truly biological zeros, we identify differentially expressed genes in the low-coverage data and display the same genes in the high-coverage data. Figure 4 shows that the majority of the zeros found by the model stay zero even when many more reads are sequenced, suggesting that the model is successful in distinguishing biological versus technical zeros.

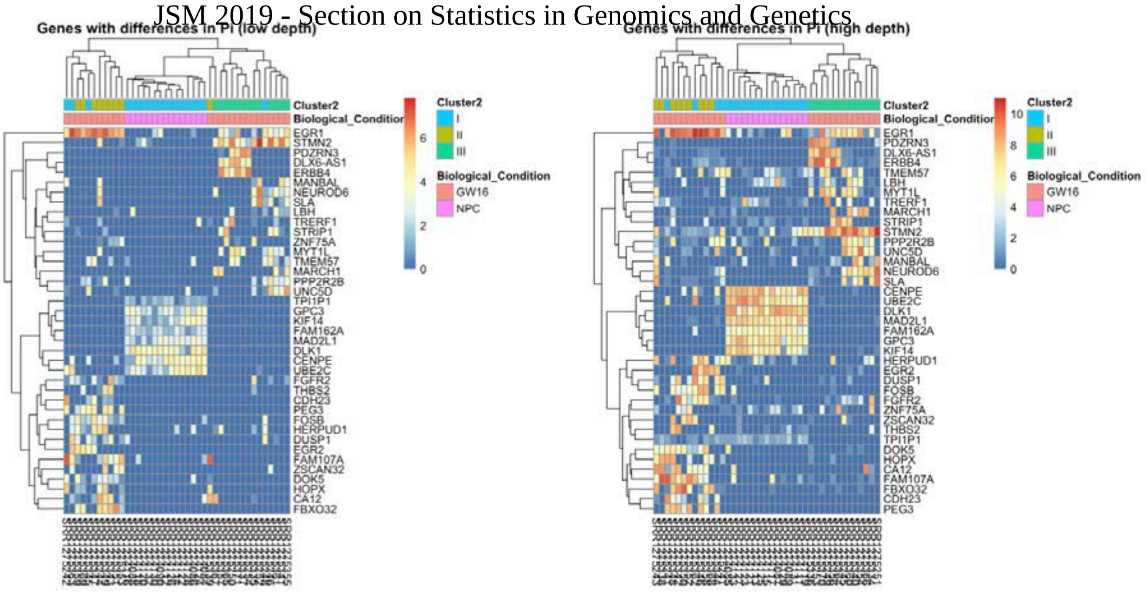

Figure 4: Differentially expressed genes identified in the low-coverage Fluidigm data (left panel) and displayed in the high-coverage Fluidigm data (right panel).

4.4 Computational Considerations

Since the ZINB model is a flexible generalization of the negative binomial, one could argue that the former should be preferable, since it would be able to fit the data even in the absence of zero inflation. However, in such cases, the extra zeros may cause an increase in variability that might be captured either by an increase in the estimated dispersion parameter or mixing parameter, making the estimation procedure unstable. Moreover, the computational burden of estimating almost twice as many parameters with respect to the negative binomial model makes the model unappealing for large datasets. The negative binomial is three times faster than the ZINB for a dataset with 1000 cells and is more scalable, being linear in the number of cells.

5. Discussion

In this paper, we focus on the effects of explicitly modeling zero inflation in single-cell RNA-seq data, with particular focus on the problem of dimensionality reduction and differential expression. While UMI data contain at least as many zero counts as non-UMI data, it has been shown that, due to a decrease in mean, these zeros can be explained by a negative binomial distribution, without any special care.

Here, we show that a negative binomial factor analysis model is sufficient to capture the biologically relevant signal in the data and is much faster than its zero-inflated counterpart. However, explicitly modeling the difference between biological and technical zeros may be beneficial in the context of differential expression, in which interesting genes might show a differential pattern of “on / off” expression in the compared conditions.

The negative binomial factor analysis model is implemented in the *zinbwave* Bioconductor package, starting from version 1.7.4, and can be used by specifying the option `zeroinflation=TRUE` in the `zinbwave` function.

Acknowledgements

The author thanks Sandrine Dudoit, Hector Roux de Bézieux, Philippe Boileau, and Kelly Street for constructive discussions on the presence of zero inflation in real single cell datasets, Monica Chiogna for her helpful comments on differential expression, and Clara Bagatin for her contribution to the zinbwave package.

References

Anders, Simon, and Wolfgang Huber. 2010. "Differential expression analysis for sequence count data". *Genome Biology* 11 (10): R106.

Bullard, James H, et al. 2010. "Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments". *BMC Bioinformatics* 11 (1): 94.

Cole, Michael B, et al. 2019. "Performance assessment and selection of normalization procedures for single-cell RNA-seq". *Cell Systems* 8 (4): 315–328.

Finak, Greg, et al. 2015. "MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data". *Genome Biology* 16 (1): 278.

Hansen, Kasper D., Davide Risso, and Stephanie Hicks. 2019. *TENxPBMCData: PBMC data from 10X Genomics*. R package version 1.3.0.

Islam, Saiful, et al. 2014. "Quantitative single-cell RNA-seq with unique molecular identifiers". *Nature Methods* 11 (2): 163.

Kharchenko, Peter V, Lev Silberstein, and David T Scadden. 2014. "Bayesian approach to single-cell differential expression analysis". *Nature Methods* 11 (7): 740.

Marinov, Georgi K, et al. 2014. "From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing". *Genome Research* 24 (3): 496–510.

Marioni, John C, et al. 2008. "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays". *Genome Research* 18 (9): 1509–1517.

McCarthy, Davis J., et al. 2017. "Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R". *Bioinformatics* 33 (8): 1179–1186.

McCarthy, Davis J, Yunshun Chen, and Gordon K Smyth. 2012. "Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation". *Nucleic Acids Research* 40 (10): 4288–4297.

Pierson, Emma, and Christopher Yau. 2015. "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis". *Genome Biology* 16 (1): 241.

Pollen, Alex A, et al. 2014. "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex". *Nature Biotechnology* 32 (10): 1053.

Risso, Davide, Michael Cole, and Aaron Lun. 2019. *scRNAseq: Collection of Public Single-Cell RNA-Seq Datasets*. R package version 1.99.2.

Risso, Davide, et al. 2018. "A general and flexible method for signal extraction from single-cell RNA-seq data". *Nature Communications* 9 (1): 284.

Robinson, Mark D, Davis J McCarthy, and Gordon K Smyth. 2010. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data". *Bioinformatics* 26 (1): 139–140.

Stan Development Team. 2019. *RStan: the R interface to Stan*. R package version 2.19.2.

Svensson, Valentine. 2019. “Droplet scRNA-seq is not zero-inflated”. *bioRxiv*: 582064.

Townes, F William, et al. 2019. “Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model”. *bioRxiv*: 574574.

Van De Wiel, Mark A, et al. 2013. “Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors”. *Biostatistics* 14 (1): 113–128.

Wagner, Allon, Aviv Regev, and Nir Yosef. 2016. “Revealing the vectors of cellular identity with single-cell genomics”. *Nature Biotechnology* 34 (11): 1145.