Dimensionality reduction for zero-inflated single cell gene expression analysis

Emma Pierson[1] and Christopher Yau[1,2,*]

[1]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford, OX1 3TG. [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN.

*Corresponding author e-mail: cyau@well.ox.ac.uk

Abstract: Single cell RNA-seq data allows insight into normal cellular function and diseases including cancer through the molecular characterisation of cellular state at the single-cell level. Dimensionality reduction of such high-dimensional datasets is essential for visualization and analysis, but single-cell RNA-seq data is challenging for classical dimensionality reduction methods because of the prevalence of dropout events leading to zero-inflated data. Here we develop a dimensionality reduction method, (Z)ero (I)nflated (F)actor (A)nalysis (ZIFA), which explicitly models the dropout characteristics, and show that it improves performance on simulated and biological datasets.

Text: Single cell RNA expression analysis (scRNA-seq) is revolutionizing whole-organism science[1] allowing the unbiased identification of previously uncharacterized molecular heterogeneity at the cellular level. Statistical analysis of single cell gene expression profiles can highlight putative cellular subtypes, delineating subgroups of T-cells[2], lung cells[3] and myoblasts[4]. These subgroups can be clinically relevant: for example, individual brain tumors contain cells from multiple types of brain cancers, and greater tumor heterogeneity is associated with worse prognosis[5].

Despite the success of early single cell studies, the statistical tools that have been applied to date are largely generic, rarely taking into account the particular structural features of single cell expression data. In particular, single cell gene expression data contains an abundance of dropout events that lead to zero expression measurements. These dropout events may be the result of technical sampling effects (due to low transcript numbers) or real biology arising from stochastic transcriptional activity (**Fig. 1a**). Here, we show that the performance of standard dimensionality-reduction algorithms on high-dimensional, single cell expression data can be perturbed by the presence of zero-inflation making them sub-optimal. We present a new dimensionality-reduction model, **Z**ero-**I**nflated **F**actor **A**nalysis (ZIFA), that explicitly accounts for the presence of dropouts, and demonstrate that ZIFA outperforms other methods on simulated data and single cell data from recent scRNA-seq studies[2,4,6,7].

The fundamental empirical observation that underlies the zero-inflation model in ZIFA is that the dropout rate for a gene depends on the expected expression level of that gene in the population. Genes with lower expression magnitude are more likely to be affected by dropout than genes that are expressed with greater magnitude. In particular, if the mean level of non-zero expression is given by μ and the dropout rate for that gene by p_0, we have found that this dropout relationship can be approximately modelled with a parametric form $p_0 = \exp(-\lambda\mu^2)$, where λ is a fitted parameter, based on a double exponential function. This relationship is consistent with previous investigations[8,14] and holds in many existing single cell datasets (**Fig. 1b**). The use of this parametric form permits fast, tractable linear algebra computations in ZIFA enabling its use on realistically sized datasets in a multivariate setting.

ZIFA adopts a latent variable model based on the Factor Analysis (FA) framework[9] and augments it with an additional zero-inflation modulation layer. Like FA, the data generation process assumes that the

separable cell states or sub-types initially exist as points in a latent (unobserved) low-dimensional space. These are then projected onto points in a latent high-dimensional gene expression space via a linear transformation and the addition of Gaussian-distributed measurement noise. Each measurement then has some probability of being set to zero via the dropout model that *modulates* the latent distribution of expression values. This allows us to account for observed zero-inflated single cell gene expression data (**Fig. 1c**). The scaling parameter in the dropout model can allow for a large range of dropout-expression profiles (**Fig. 1d**). ZIFA performs maximum likelihood-based statistical inference assuming this data generative model to infer the latent, low-dimensional cell state representation from the observed zero-inflated high-dimensional gene expression data. As with FA we were able to derive a fast and scalable expectation-maximization (EM) algorithm[10] to fit the model. The algorithm structurally resembles the equivalent EM algorithm for FA but incorporates additional data imputation steps to estimated expected gene expression levels for observed null values.

We tested the relative performance of ZIFA against Principal Components Analysis (PCA), Probabilistic PCA (PPCA), Factor Aanalysis and, where appropriate, non-linear techniques including Stochastic Neighbour Embedding (t-SNE)[11], Isomap[12], and Multidimensional Scaling[13] (MDS). First, we generated simulated datasets according to the PPCA/FA data generative model with the addition of one of three dropout models (i) a double exponential model (as assumed by ZIFA), (ii) a linear decay model and (iii) a missing-at-random uniform model. The latter two models were designed to test the robustness of ZIFA to extreme misspecification of the dropout model. Data was simulated under a range of different conditions by varying noise levels, dropout rates, number of latent dimensions and number of genes.

We applied the dimensionality reduction methods to the simulated data sets and obtained the output of each algorithm as a projection of the observed zero-inflated data on to a lower-dimensional latent space (**Fig. 2a**). We defined performance as the Spearman correlation between the distances of the projected data in the inferred latent low-dimensional space versus the distances in the actual latent space used in the simulation. Overall, ZIFA outperformed standard dimensionality reduction algorithms under all simulations. This occurred regardless of whether zeros were added following the assumed decaying squared exponential model (**Fig. 2b**), a linear model (**Supplementary Fig. 1B**), or missing-at-random model (**Supplementary Fig. 1C**). Although the data sets was generated according to a PPCA/FA model (up to the dropout stage), in the presence of cells with genes possessing zero expression, PPCA and FA will be sub-optimal compared to ZIFA. We provide an interactive illustration of our model[15], showing how it more clearly resolves separable latent clusters from zero-inflated data.

We next sought to test these methods in an experiment based on real single cell expression datasets[2,4,6,7]. In this case, the "true" latent space is unknown and we are unable to measure performance as with the previous simulated data experiment. Instead, for each of the data sets, we took random subsets of 25, 100, 250 and 1,000 genes and applied ZIFA, PPCA and FA to each subset assuming 5 latent dimensions. We then compared, for each gene in the subset, the predicted data distribution obtained by each dimensionality reduction method to the empirical data distribution by computing the divergence between these two distributions (see Supplementary Information). Using this criterion we found that predictive distributions from PPCA and FA showed high divergence for genes that exhibited a high dropout rate or possessed a low non-zero expression level. This meant that the predictive data distributions were a poor fit for the empirical data. ZIFA performance was largely unaffected in contrast (**Fig. 2c**). Example predictive model fits are shown for the T-cell data set[2] for three genes: *Plscr3*, *Ulk2* and *Ncrna00085* (**Fig. 2c**). The statistical frameworks underlying PPCA and FA employ Gaussianity assumptions that are unable to explicitly account for zero-inflation in single cell expression data. The dropout model used by ZIFA modulates this Gaussianity assumption allowing for zero-inflation leading to

drastically improved modelling accuracy. Across the four data sets we found that the predictive distribution derived by ZIFA was superior to those of PPCA and FA on *at least* 80% of the genes examined and often over 95% (**Supplementary Table 1**).

The density of dropout events in scRNA-seq data can render classical dimensionality-reduction algorithms unsuitable and to-date it has not been possible to assess the potential ramifications of applying such methods on zero-inflated data. We have modified the PPCA/FA framework to account for dropout to a produce a "safe" method for visualization and clustering of single-cell gene expression data that provides robustness against such uncertainties. Our methodology differs from approaches, such as Robust PCA, that model corrupted observations. ZIFA treats dropouts as observations, not outliers, whose occurrence properties have been characterised using an empirically informed statistical model. ZIFA is also potentially applicable to other zero-inflated data where there is a negative correlation between the frequency with which a measurement feature is zero and its mean signal magnitude in non-zero samples. A Python-based software implementation is available online: https://github.com/epierson9/ZIFA.

Methods

Detailed simulation and analytical methodology as well as mathematical derivations are contained in the Supplementary Information.

Acknowledgements

E.P. acknowledges support from the Rhodes Trust. C.Y. is supported by a UK Medical Research Council New Investigator Research Grant (Ref. No. MR/L001411/1), the Wellcome Trust Core Award Grant Number 090532/Z/09/Z, the John Fell Oxford University Press (OUP) Research Fund and the Li Ka Shing Foundation via a Oxford-Stanford Big Data in Human Health Seed Grant.

Author Contributions

E.P. and C.Y. conceived the study and developed the algorithms. E.P. performed data analysis and developed the software implementation. E.P. and C. Y. wrote the manuscript.

Competing Financial Interests

The authors declare no competing financial interests.

References

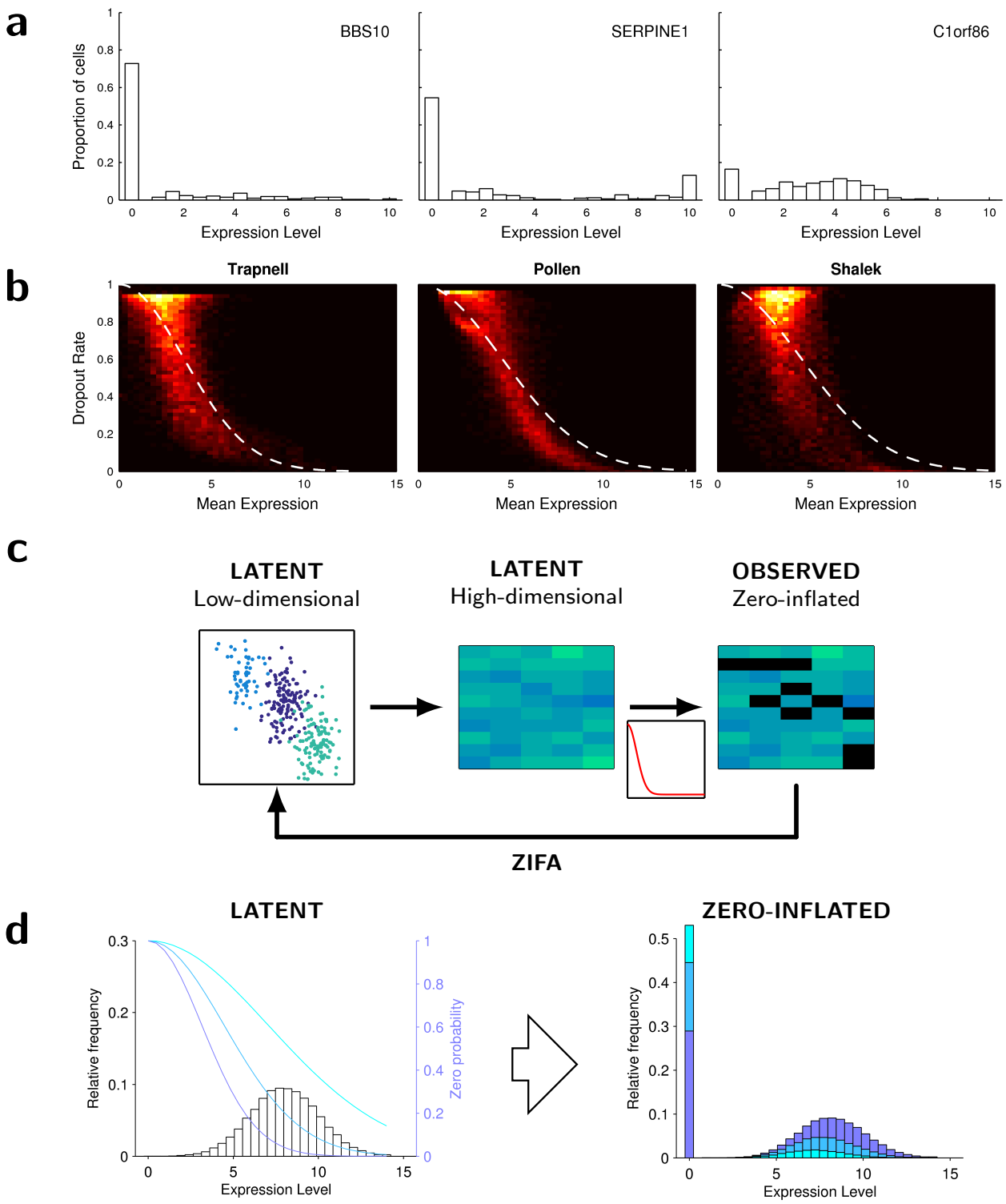
1. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.* **14,** 618–30 (2013).
2. Buettner, F. *et al.* Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.* **advance on,** (2015).
3. Treutlein, B. *et al.* Reconstructing lineage hierarchies of the distal lung epithelium using single-cell RNA-seq. *Nature* **509,** 371–5 (2014).
4. Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.* **31,** 46–53 (2012).
5. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-.).* **344,** 1396–1401 (2014).
6. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* (2014).

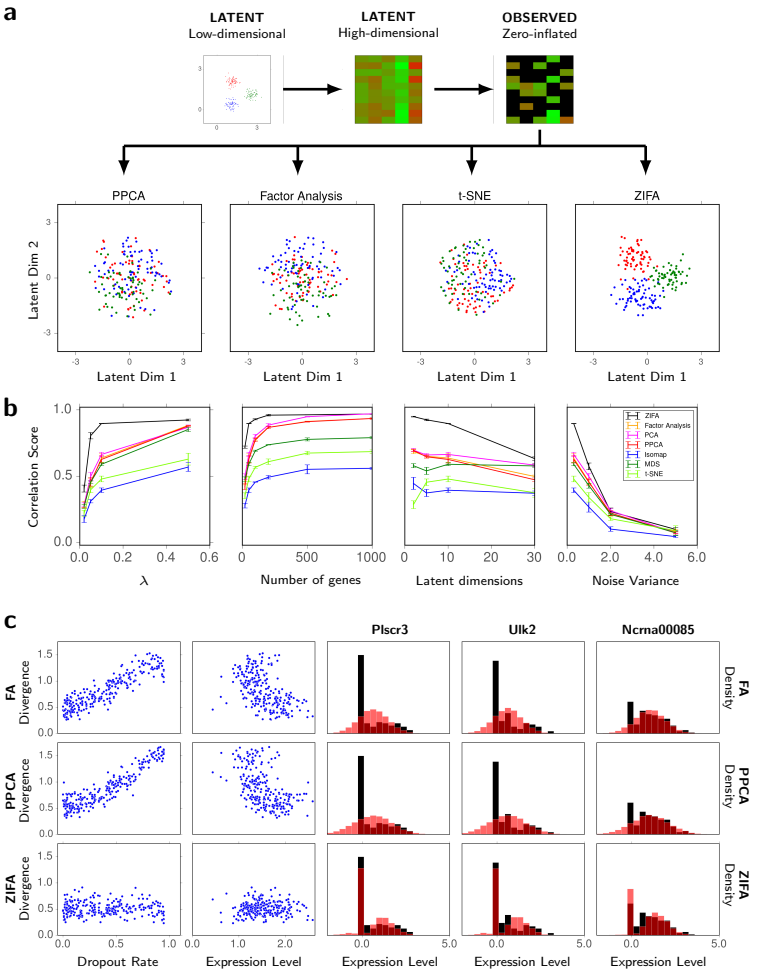
7. Pollen, A. A. *et al.* Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex. *Nat. Biotechnol.* (2014). doi:10.1038/nbt.2967
8. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nat. Biotechnol.* **30,** 777–782 (2012).
9. Tipping, M. E. & Bishop, C. M. Probabilistic Principal Component Analysis. *J. R. Stat. Soc. Ser. B (Statistical Methodol.* **61,** 611–622 (1999).
10. Dempster, A., Laird, N. & Rubin, D. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. …* (1977).
11. Maaten, L. Van der & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. …* (2008).
12. Balasubramanian, M. & Schwartz, E. The isomap algorithm and topological stability. *Science (80-.).* (2002).
13. Kruskal, J. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* (1964).
14. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Meth.* **11**, 740-742 (2014).
15. Pierson, E. Yau, C. ZIFA Demo. http://cs.stanford.edu/~emmap1/visualizations/ZIFA/demo.html.

List of Figures

Figure 1: Zero-inflation in single cell expression data. (a) Illustrative distribution of expression levels for three randomly chosen genes shows an abundance of single cells exhibiting null expression[7]. (b) Heatmaps showing the relationship between dropout rate and mean non-zero expression level for three published single cell data sets[2,4,6] including an approximate double exponential model fit. (c) Flow diagram illustrating the data generative process used by ZIFA. (d) Illustrative plot showing how different values of λ in the dropout-mean expression relationship (blue lines) can modulate the latent gene expression distribution to give a range of observed zero-inflated data.

Figure 2: Performance comparison of dimensionality reduction techniques. (a) Toy simulated data example illustrating the performance of ZIFA compared to standard dimensionality reduction algorithms. (b) Performance on simulated datasets based on correlation score between the estimated and true latent distances as a function of λ (larger λ, lower dropout rate), number of genes and latent dimensions and noise level used in the simulations. (c) Plots showing the divergence between the predictive and empirical data distributions as a function of dropout rate and mean expression level for FA, PPCA and ZIFA. Illustrative predictive performance and model fits (red) on the T-cell single cell data set (black)[2].


a
BBS10
SERPINE1
C1orf86
Proportion of cells
Expression Level
b
Trapnell
Pollen
Shalek
Dropout Rate
Mean Expression
c
LATENT
Low-dimensional
LATENT
High-dimensional
OBSERVED
Zero-inflated
ZIFA
d
LATENT
ZERO-INFLATED
Relative frequency
Zero probability
Expression Level


a
LATENT
Low-dimensional
LATENT
High-dimensional
OBSERVED
Zero-inflated
PPCA
Factor Analysis
t-SNE
ZIFA
Latent Dim 2
Latent Dim 1
b
Correlation Score
λ
Number of genes
Latent dimensions
Noise Variance
ZIFA
Factor Analysis
PCA
PPCA
Isomap
MDS
t-SNE
c
Plscr3
Ulk2
Ncrna00085
FA
PPCA
ZIFA
Divergence
Density
Dropout Rate
Expression Level