Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis

Ghislain Durif, Laurent Modolo, Jeff Mold, Sophie Lambert-Lacroix, Franck Picard

Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis

G. Durif[1,2], L. Modolo[1,3,4], J. E. Mold[4], S. Lambert-Lacroix[5] and F. Picard[1]

[1] LBBE, UMR CNRS 5558, Université Lyon 1, F-69622 Villeurbanne, France
[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, F-38000 Grenoble, France
[3] LBMC UMR 5239 CNRS/ENS Lyon, F-69007 Lyon, France
[4] Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden
[5] UMR 5525 Université Grenoble Alpes/CNRS/TIMC-IMAG, F-38041 Grenoble, France ghislain.durif@inria.fr

The combination of massive parallel sequencing with high-throughput cell biology technologies has given rise to single-cell Genomics. Similar to the paradigm shift of the 90s characterized by the first molecular profiles of tissues, it is now possible to characterize molecular heterogeneities at the cellular level (Saliba et al., 2014). The statistical characterization of heterogeneities in single-cell expression data thus requires an appropriate model, since the transcripts abundance is quantified for each cell using read counts. Hence, standard methods based on Gaussian assumptions are likely to fail to catch the biological variability of lowly expressed genes, and Poisson or Negative Binomial distributions constitute an appropriate framework (Chen et al., 2016). Moreover, dropouts, either technical (due to sampling difficulties) or biological (no expression or stochastic transcriptional activity), constitute another major source of variability in scRNA-seq (single-cell RNA-seq) data, which has motivated the development of the so-called Zero-Inflated models (Kharchenko et al., 2014). A standard and popular way of quantifying and visualizing the variability within a dataset is dimension reduction, principal component analysis (PCA) being the most widely used technique in practice. Model-based PCA (Collins et al., 2001) offers the unique advantage to be adapted to the data distribution and to be based on an appropriate metric, the Bregman divergence. It consists in specifying the distribution of the data through a statistical model. A probabilistic zero-inflated version of the Gaussian PCA was proposed by Pierson & Yau (2015) in the context of single cell data analysis (the ZIFA method). However, scRNA-seq data may be better analyzed by methods dedicated to count data such as the Non-negative Matrix Factorization (Lee & Seung, 1999, NMF) or the Gamma-Poisson factor model (Cemgil, 2009). However, none of the currently available dimension reduction methods fully model single-cell expression data, characterized by overdispered zero inflated counts (Zappia et al., 2017). Our method is based on a probabilistic count matrix factorization (pCMF). We propose a dimension reduction method that is dedicated to over-dispersed counts with dropouts, in high dimension. Our factor model takes advantage of the Poisson Gamma representation to model counts from scRNA-seq data (Zappia et al., 2017). In particular, we use Gamma priors on the distribution of principal components. We model

dropouts with a Zero-Inflated Poisson distribution, and we introduce sparsity in the model thanks to a spike-and-slab approach (Malsiner-Walli & Wagner, 2011) that is based on a two component sparsity-inducing prior on loadings (Titsias & Lázaro-Gredilla, 2011). The model is inferred using a variational EM algorithm that scales favorably to data dimension, as compared with Markov Chain Monte Carlo (MCMC) methods (Blei et al., 2017). Then we propose a new criterion to assess the quality of fit of the model to the data, as a percentage of explained deviance, because the standard variance reduction that is used in PCA needs to be adapted to the new framework dedicated to counts. We show that pCMF better catches the variability of simulated data and experimental scRNA-seq datasets. Finally, pCMF is available in the form of a R package available at https://gitlab.inria.fr/gdurif/pCMF.

References

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, (just-accepted).

Cemgil, A. T. (2009). Bayesian inference for nonnegative matrix factorisation models. *Computational Intelligence and Neuroscience*, *2009*.

Chen, H.-I. H., Jin, Y., Huang, Y., & Chen, Y. (2016). Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics*, *17*(Suppl 7).

Collins, M., Dasgupta, S., & Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. In *Advances in Neural Information Processing Systems*, (pp. 617–624).

Kharchenko, P. V., Silberstein, L., & Scadden, D. T. (2014). Bayesian approach to single-cell differential expression analysis. *Nature Methods*, *11*(7), 740.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Malsiner-Walli, G., & Wagner, H. (2011). Comparing spike and slab priors for Bayesian variable selection. *Austrian Journal of Statistics*, *40*(4), 241–264.

Pierson, E., & Yau, C. (2015). ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biology*, *16*, 241.

Saliba, A.-E., Westermann, A. J., Gorski, S. A., & Vogel, J. (2014). Single-cell RNA-seq: Advances and future challenges. *Nucleic Acids Research*, *42*(14), 8845–8860.

Titsias, M. K., & Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multi-task and multiple kernel learning. In *Advances in Neural Information Processing Systems*, (pp. 2339–2347).

Zappia, L., Phipson, B., & Oshlack, A. (2017). Splatter: Simulation of single-cell RNA sequencing data. *Genome Biology*, *18*, 174.