Variational Inference in Probabilistic Single-cell RNA-seq Models *

Pedro F. Ferreira[1 [0000−0003−0559−6125]], Alexandra M. Carvalho[1,2 [0000−0001−6607−7711]], and Susana Vinga[1,3 [0000−0002−1954−5487]]

[1] Instituto Superior Técnico, ULisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[2] Instituto de Telecomunicações, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
[3] INESC-ID, R. Alves Redol 9, 1000-029 Lisboa, Portugal
{pedro.fale,alexandra.carvalho,susanavinga}@tecnico.ulisboa.pt

Abstract. Single-cell sequencing technology holds the promise of unravelling cell heterogeneities hidden in ubiquitous bulk-level analyses. However, limitations of current experimental methods also pose new obstacles that prevent accurate conclusions from being drawn. To overcome this, researchers have developed computational methods which aim at extracting the biological signal of interest from the noisy observations. In this paper we focus on probabilistic models designed for this task. Particularly, we describe how variational inference constitutes a powerful inference mechanism for different sample sizes, and critically review two recent scRNA-seq models which use it.

Keywords: scRNA-seq · probabilistic modelling · Bayesian inference · dimensionality reduction · imputation

1 Scientific Background

Single-cell RNA-sequencing (scRNA-seq) has emerged in the last decade as a key technology in using gene expression to study cell heterogeneity [1]. With the obtained data, researchers can, for example, apply clustering algorithms to identify cell types and find genes which are differentially expressed between two conditions.

In scRNA-seq data, each observation is a cell and, for each cell, the expression of all detected genes is measured through the set of all RNA molecules present, i.e., its transcriptome. Specifically, each entry in the $N \times P$ data matrix, where N is the number of cells and P the number of genes, contains the number of mRNA molecules corresponding to gene p detected in cell n. Depending on the experimental protocol and quality control pipelines, data set sizes may vary from

*Supported by the EU Horizon 2020 research and innovation program (grant No. 633974 – SOUND project), and the Portuguese Foundation for Science & Technology (FCT), through UID/EMS/50022/2019 (IDMEC,LAETA), UID/EEA/50008/2019 (IT), UID/CEC/50021/2019 (INESC-ID), PTDC/EMS-SIS/0642/2014, PTDC/CCI-CIF/29877/2017, PTDC/EEI-SII/1937/2014, IF/00653/2012, and by internal IT projects QBigData and RAPID.

hundreds to millions of cells [1] and from hundreds to tens of thousands of genes sequenced.

Although increasingly available, scRNA-seq data suffer from multiple confounding factors which may hide the biological signal of interest from analysis. These include varying sequencing depths and mRNA capture efficiency which lead to zero-inflated observations and library size (total number of mRNA molecules detected per cell) dispersion, as well as batch effects [1]. Because of this, applying generic computational methods for further downstream analyses, such as dimensionality reduction, clustering, or differential expression, yields spurious results. Indeed, while PCA may capture the existence of different clusters of cells in a data set, its principal components are highly correlated with technical factors [2]. Researchers have thus developed methods to extract only the biological signal of interest; the most commonly tackled issue is the unrealistic abundance of zero counts, termed "dropouts".

In this report, we focus on methods based on probabilistic models of scRNA-seq. Probabilistic modelling [3] constitutes a powerful framework for the disentanglement of multiple factors of variation in the stochastic generative process underlying the data. In general, probabilistic scRNA-seq models assume a lower-dimensional representation of the data, which is mapped into the observation space by some transformation, allowing for dimensionality reduction and a dropout-inducing process. This framework is particularly powerful because, by explicitly accounting for the different assumed factors of variation, it can be used for multiple downstream tasks after fitting to the data.

Probabilistic models define a joint probability distribution $p(\mathbf{X}, \mathbf{Z})$ over the observed data $\mathbf{X}$ and a set of latent variables $\mathbf{Z}$. These encode structure in the data, via some prior distribution $p(\mathbf{Z})$, and are related to the observations via the likelihood distribution $p(\mathbf{X}|\mathbf{Z})$. After defining such a model, inference of the latent variables is made via their probability distribution conditioned on the data, $p(\mathbf{Z}|\mathbf{X})$. This is called the posterior probability distribution and, according to Bayes' theorem, is given by

$$p(\mathbf{Z}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{Z})p(\mathbf{Z})}{p(\mathbf{X})}. \tag{1}$$

However, in general, for complex models, Eq. (1) can not be computed analytically. Approximating the posterior is thus the main computational challenge in probabilistic modelling. The common approach for this task is to use Markov Chain Monte Carlo sampling methods, with Gibbs sampling being the gold standard [3]. However, variational inference techniques are generally able to provide similar performance at a possibly lower computational cost, making them more suitable for large data sets.

In the following sections we describe two recent probabilistic models designed for scRNA-seq data which use variational inference to infer their latent variables: *Probabilistic Count Matrix Factorization* (pCMF) [4] and *Single Cell Variational Inference* (scVI) [5]. While both use the same inference engine, the modelling details of each allow for different techniques to be used, which we outline. The

performance of both models is measured in terms of cell type separability in the lower-dimensional latent space, and dropout imputation error. As a baseline for comparison, we consider a state-of-the-art scRNA-seq probabilistic model, *Zero Inflated Factor Analysis* (ZIFA) [6].

2 Materials and Methods

We first describe variational inference. Then we describe pCMF and scVI. In particular, we aim at illustrating how the variational scheme allows for efficient inference of complex probabilistic models, both in small N, large P data sets, and the inverse.

2.1 Variational inference

In variational inference the true posterior $p(\mathbf{Z}|\mathbf{X})$ defined in Eq. (1) is approximated via a distribution $q(\mathbf{Z};\mu)$, which belongs to a certain family $\mathcal{Q}$, over the latent variables $\mathbf{Z}$ with free parameters μ [7]. These parameters are adjusted so as to minimize some distance between $q(\mathbf{Z})$ and $p(\mathbf{Z}|\mathbf{X})$. We thus turn inference into an optimization problem. The most commonly used distance metric between these distributions is the Kullback-Leibler (KL) divergence. In this case, the optimization problem becomes

$$q(\mathbf{Z}) = \underset{q(\mathbf{Z})\in\mathcal{Q}}{\operatorname{argmin}} \operatorname{KL}(q(\mathbf{Z}) \parallel p(\mathbf{Z}|\mathbf{X})). \tag{2}$$

The objective in Eq. (2) is not available because it depends on the posterior distribution which we aim at approximating. However, we can re-write the KL divergence in terms of a lower bound of $p(\mathbf{X})$ which we call the Evidence Lower BOund (ELBO). Minimizing the KL divergence is now achieved by maximizing the ELBO:

$$q(\mathbf{Z}) = \underset{q(\mathbf{Z})\in\mathcal{Q}}{\operatorname{argmax}} \operatorname{ELBO}(\mu) = \mathrm{E}_q\left[\log(p(\mathbf{X},\mathbf{Z}))\right] - \mathrm{E}_q\left[\log(q(\mathbf{Z};\mu))\right]. \tag{3}$$

This optimization is constrained not only by the family of distributions $\mathcal{Q}$ we choose, but also by the widely used mean-field approximation, where we assume each of the M latent variables to be independent from all the others and governed by their own variational density [7]. This makes the ELBO a non-convex function.

The most commonly used algorithm to find the μ that correspond to a local maximum of the ELBO is coordinate ascent, which we refer in the following sections as CAVI (Coordinate Ascent Variational Inference). CAVI algorithms can be easily derived for conditionally conjugate models. More recently, ELBO optimization has been generalized into the wider class of non-conditionally conjugate models, effectively allowing the design of more expressive models [7].

2.2 Probabilistic Count Matrix Factorization (pCMF)

This model consists of a Bayesian matrix factorization method for count data. Its latent variables are $\mathbf{U}$, $\mathbf{D}$ and $\mathbf{V}$. [4] $\mathbf{U}$ represents the cells in a lower-dimensional space of size $K < P$, $\mathbf{V}$ is the map from $\mathbf{U}$ to the observation space, and $\mathbf{D}$ models the occurrence of dropout in each observation.

By considering Gamma priors on $\mathbf{U}$ and $\mathbf{V}$, pCMF models the over-dispersion of the count data. $\mathbf{D}$ is given by a Bernoulli distribution. The model is represented graphically in Fig. 1 and defined by the generative process which Algorithm 1 outlines. [5]

Algorithm 1 Generative process for pCMF

For each cell n:
  For each k, sample a latent factor:
  $U_{nk} \sim \text{Gamma}\left(\alpha_{k1}, \alpha_{k2}\right)$.
For each gene p:
  For each k, sample a factor load:
  $V_{pk} \sim \text{Gamma}\left(\beta_{k1}, \beta_{k2}\right)$.
For each cell n and gene p,
  Sample dropout event:
  $D_{np} \sim \text{Bernoulli}\left(\pi_p\right)$.
  Sample the observed count:
  $X_{np} \sim \text{Poisson}\left(\left(1 - D_{np}\right)\mathbf{U}_n\mathbf{V}_p^T\right)$.

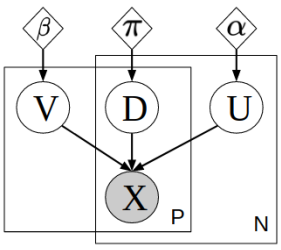


Fig. 1: Graphical representation of pCMF.

Because this model is conditionally conjugate (if we consider an auxiliary variable, see [4] for details), the posterior can be approximated via a CAVI algorithm.

While in traditional machine learning settings there are more observations than features, and thus an un-regularized point estimate of the global variables (commonly called "parameters") is enough, most initial scRNA-seq data sets, and the ones pCMF was designed for, are such that $P \gg N$. In these cases, the number of global variables (V_{pk}) is larger than the number of local variables (U_{nk}) in the model, which makes correct estimation of the global variables difficult. Thus the need for a Bayesian approach even for the global variables in this case.

However, CAVI requires the whole data set to compute each variational parameter update. As such, inference of pCMF on a data set with a much larger number of cells would imply a great computational effort, due to the need to infer a posterior over the global variables. In that case, it would suffice to use point estimates for the global variables instead, for example in an Expectation-Maximization algorithm.

[4] For brevity, here we do not consider the sparse loadings of the original model. In our experiments the resulting performance did not change significantly.

[5] $\alpha_{k1,2}$, $\beta_{k1,2}$ and π_p are fixed hyperparameters which can be estimated in an Expectation-Maximization scheme. See the original paper for details.

2.3 Single Cell Variational Inference (scVI)

scVI models the distribution of observed counts as conditioned on: **L**, the variations due to capture efficiency and sequencing depth of each cell; **W**, the normalized mean expressions; dropout events **D**; θ for gene-specific dispersion and **Z**, a lower-dimensional space where biological variability is encoded. It can also include the batch annotation of each cell in order to subtract batch effects from the biological signal.

Additionally, scVI utilizes neural networks to specify non-linear transformations between latent variables. Notably, it associates the cells' latent representations **Z** with the probability of dropout occurrence, encoded by a Bernoulli distribution on **D** whose parameter is given by a neural network f_D with output in the $[0, 1]$ interval. Another neural network f_W is used to map from **Z** to the original-dimensional space containing **W**, which is encoded by a Gamma-distributed random variable. The generative process is described in Algorithm 2 and Fig. 2 presents the corresponding graphical model. [6,7]

Algorithm 2 Generative process for scVI

For each cell n:
  For each k, sample a latent factor:
  $Z_{nk} \sim \text{Normal}(0, 1)$.
  Sample a cell-scaling factor:
  $L_n \sim \text{LogNormal}\left(l_\mu, l_\sigma^2\right)$.
For each cell n and gene p,
  Sample mean expression:
  $W_{np} \sim \text{Gamma}\left(f_W\left(Z_n\right), \theta_p\right)$.
  Sample dropout event:
  $D_{np} \sim \text{Bernoulli}\left(f_D\left(Z_n\right)\right)$.
  Sample the observed count:
  $X_{np} \sim \text{Poisson}\left(\left(1 - D_{np}\right) L_n W_{np}\right)$.

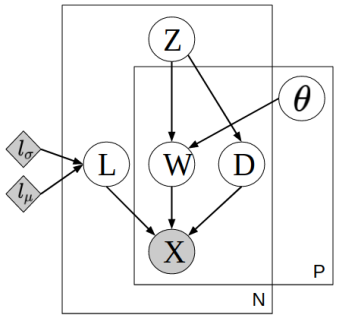


Fig. 2: Graphical representation of scVI.

Inference of scVI's latent variables is performed using neural networks specifying the approximate posterior distributions $q(\mathbf{Z})$ and $q(\mathbf{L})$ (the "inference networks" [8]), in which the other latent variables are integrated out. This allows for inference to be reduced to optimizing the weights of the four neural networks: f_W, f_D and $q(\mathbf{Z})$, $q(\mathbf{L})$ [5]. Unlike CAVI, this is a general mechanism possible for models without conditional conjugacy. It also makes inference ammenable to stochastic optimization, meaning global variables can be estimated using small subsets of the data per iteration.

While the use of neural networks allows for great model expressiveness, the typically large number of parameters to fit may render them inadequate for small

[6] In these simplified descriptions we ignore the batch annotation observations, for brevity.

[7] l_μ and l_σ^2 are the observed log-library size mean and variance, respectively.

sample sizes. In addition, the ability to approximate the true posterior is limited by the flexibility of the inference networks.

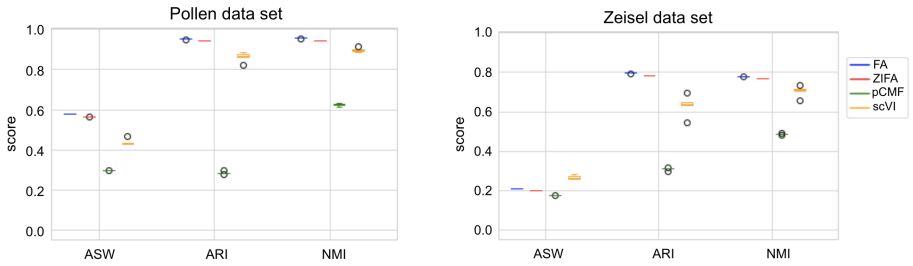
3 Results

We test the methods described in Section 2 on real scRNA-seq data. For ZIFA and scVI, we use the implementations provided by the authors with the original publications. For pCMF we used our own implementation which allows for more flexibility in the inference scheme (i.e., inclusion of sparsity and hyperparameter estimation) than the one provided by the authors. We did, however, compare our implementation with the original one, and the same results were obtained. In our tests, we set the latent space dimensionality to $K = 10$ and apply all models to two real data sets, whose main characteristics are summarized in Table 1.

Table 1: Statistics of the considered experimental data sets.

Data set	**# cells**	**# genes**	**# cell types**	**% zeros**
Pollen [9]	249	6982	11	25.33
Zeisel [10]	3005	558	7	29.01

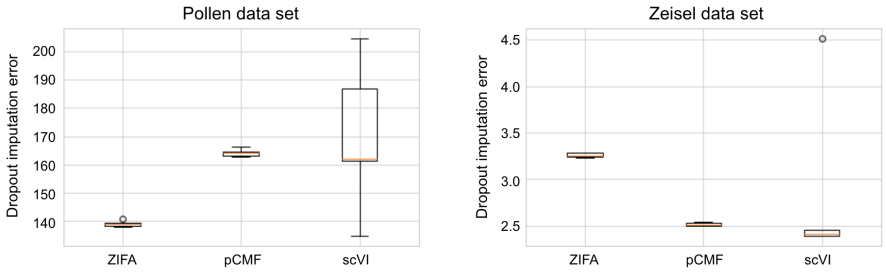
Fig. 3 shows the evaluation of cluster separability in the latent space using different metrics: Average Silhouette Width (ASW), Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). For ARI and NMI, we used the K-means clustering method to obtain partitions. Factor Analysis (FA) and ZIFA perform similarly and always better than pCMF and scVI. pCMF is always worse than scVI.



(a) (b)

Fig. 3: Boxplots of ASW, ARI and NMI scores for each method for five repetitions on the (a) Pollen and (b) Zeisel data sets. Lines indicate no variation in the score for the five runs.

Following [5] we apply dropouts to 10% of the non-zero entries in each data set and compare the values imputed by the model with the original ones. Fig. 4 shows the results for five repetitions of this process. In this sense, scVI is more sensitive than pCMF to the change of proportion between cells and genes.



(a) (b)

Fig. 4: Boxplots for the median L1 distance between imputed and original values for each scRNA-seq model fitted to five corrupted versions of the (a) Pollen and (b) Zeisel data sets.

Finally, we aim at understanding the effect of cell-specific scalings in the performance of scVI – its flexibility allows us to easily include them or not in the inferred model. Table 2 shows the results for 5-fold cross-validation (CV) on the Zeisel data set regarding imputation error and likelihood of held-out data. The results show that including cell-specific scalings to account for capture efficiency and sequencing depth improves the model fitness of scVI.

Table 2: Mean and standard deviation of scVI fitness metrics for a 5-fold CV on the Zeisel data set.

Scalings	**Dropout imputation error**	**Predictive log-likelihood**
No	4.224 (0.433)	-17245.777 (4750.590)
Yes	**2.338 (0.105)**	**-1515.624 (244.203)**

4 Conclusion

The results show that the separability of clusters in the latent space achieved by pCMF and scVI are not as good as the ones achieved by ZIFA or even FA. However, while pCMF and scVI may not provide better separations than FA and ZIFA, the explicit modelling of confounding factors guarantees that the structure they infer in the latent space is more related with actual biology rather

than technical variability. In this light, scVI is more powerful than pCMF: not only does it account for more factors of variation, but it also achieves better cluster separability.

The expressiveness of scVI also allows for better imputation of dropouts in data sets with more cells than genes, such as Zeisel's. In the reverse case, due to scVI's complexity, it is expected to underfit the data, resulting in higher imputation errors (a behaviour also observed in [5] for a different data set). In this case, additional gene filtering must be used before applying scVI.

One of the main modelling issues that allow the good results of scVI is the use of cell-specific scalings. As shown in Table 2, their inclusion results in a large increase in dropout imputation error and the likelihood assigned to held-out data.

Additionally, leveraging the modelling power of scVI is easily done via the flexible inference process based on inference networks. This proves the versatility of this recent variational inference technique, which ultimately allows for models to be designed without worrying about the inference process, thus allowing the model designer to iterate faster over different model choices.

Acknowledgements

The authors thank Ghislain Durif for the helpful discussions about pCMF.

References

1. Aleksandra A. Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C. Marioni, Sarah A. Teichmann. "The Technology and Biology of Single-Cell RNA Sequencing". *Molecular Cell*, vol.58, no.4, pp. 610-620, 2015.
2. Stephanie C. Hicks et al. "Missing data and technical variability in single-cell RNA-sequencing experiments". *Biostatistics*, 2017.
3. K. Murphy. "Machine Learning: A Probabilistic Approach". MIT Press, 2012.
4. G. Durif, L. Modolo, J. E. Mold, S. Lambert-Lacroix, F. Picard. "Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis". *arXiv*, 2018.
5. R. Lopez, J. Regier, M. B. Cole, M. Jordan, N. Yosef. "Bayesian Inference for a Generative Model of Transcriptome Profiles from Single-cell RNA Sequencing". *bioRxiv*, 2018.
6. E. Pierson and C. Yau. "ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis". *Genome Biology*, vol.16, no.1, pp. 241, 2015.
7. David M. Blei, Alp Kucukelbir and Jon D. McAuliffe. "Variational inference: A review for statisticians". *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.
8. D. Kingma and M. Welling. "Stochastic gradient VB and the variational auto-encoder". *Second International Conference on Learning Representations, ICLR*, 2014.
9. Alex A. Pollen et al. "Low-coverage single-cell mRNA sequencing reveals cellular heterogeneity and activated signaling pathways in developing cerebral cortex". *Nature Biotechnology*, vol.32, pp. 1053-1058, 2014.
10. A. Zeisel et al. "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq". *Science*, vol. 347, issue 6226, pp. 1138-1142, 2015.