Probabilistic Count Matrix Factorization for Single Cell Expression Data Analysis

Ghislain Durif, Laurent Modolo, Jeff E 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]

November 2, 2017

[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.
*Corresponding author: ghislain.durif@inria.fr

Abstract

The development of high throughput single-cell technologies now allows the investigation of the genome-wide diversity of transcription. This diversity has shown two faces: the expression dynamics (gene to gene variability) can be quantified more accurately, thanks to the measurement of lowly-expressed genes. Second, the cell-to-cell variability is high, with a low proportion of cells expressing the same gene at the same time/level. Those emerging patterns appear to be very challenging from the statistical point of view, especially to represent and to provide a summarized view of single-cell expression data. PCA is one of the most powerful framework to provide a suitable representation of high dimensional datasets, by searching for new axis catching the most variability in the data. Unfortunately, classical PCA is based on Euclidean distances and projections that work poorly in presence of over-dispersed counts that show zero-inflation. We propose a probabilistic Count Matrix Factorization (pCMF) approach for single-cell expression data analysis, that relies on a sparse Gamma-Poisson factor model. This hierarchical model is inferred using a variational EM algorithm. We show how this probabilistic framework induces a geometry that is suitable for single-cell data, and produces a compression of the data that is very powerful for clustering purposes. Our method is competed to other standard representation methods like t-SNE, and we illustrate its performance for the representation of single-cell data. We especially focus on a publicly available data set, being single-cell expression profile of neural stem cells.

1 Introduction

The combination of massive parallel sequencing with high-throughput cell biology technologies has given rise to single-cell Genomics, which refer to techniques that now provide genome-wide measurements of a cell's molecular profile either based on DNA (Zong et al., 2012), RNA (Picelli et al., 2013), or chromatin (Buenrostro et al., 2015; Rotem et al., 2015). Similar to the paradigm shift of the 90s characterized by the first molecular profiles of tissues (Golub et al., 1999), it is now possible to characterize molecular heterogeneity at the cellular level. A tissue is now viewed as a population of cells of different types, and many fields have now identified intra-tissue heterogeneities, in T cells (Buettner et al., 2015), lung cells (Trapnell et al., 2014), or myeloid progenitors (Paul et al., 2015). The construction of a comprehensive atlas of human cell types is now within our reach (Wagner et al., 2016).

The statistical characterization of heterogeneities in single-cell expression data thus requires an appropriate model, since the abundance transcripts is quantified for each cell using read counts. Hence, standard model 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. 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 (Pierson & Yau, 2015).

A standard an 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. It consist in approximating the observation matrix $\mathbf{X}_{[n\times p]}$ (n cells, p genes), by a factorized matrix of reduced rank, denoted $\mathbf{U}\mathbf{V}^T$ where $\mathbf{U}_{[n\times K]}$ and $\mathbf{V}_{[p\times K]}$ represent the latent structure in the observation and variable spaces respectively. This projection onto a lower-dimensional space (of dim. K) allows to catch gene co-expression patterns and clusters of individuals. PCA is probably one of the most studied data analysis techniques, and can be viewed either geometrically or through the light of a statistical model (Collins et al., 2001; Landgraf & Lee, 2015). Model-based PCA 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 $\mathbf{X}_{[n\times p]}$ through a statistical model, and to factorize $\mathbb{E}(\mathbf{X})$ instead of $\mathbf{X}$. On the contrary, standard PCA is based on an implicit Gaussian distribution with the ℓ_2 distance as a metric (Eckart & Young, 1936). Many distributions have been considered, especially for count data such as the Non-negative Matrix Factorization (NMF) introduced in a Poisson-based framework by Lee & Seung (1999) or the Gamma-Poisson factor model (Cemgil, 2009; Févotte & Cemgil, 2009; Landgraf & Lee, 2015). However, none of the currently available dimension reduction methods fully model single-cell expression data.

Our method is based on probabilistic count matrix factorization (pCMF). We pro-

pose 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, with the use of Gamma priors on the distribution of principal components. We model dropouts with a Zero-Inflated Poisson distribution (Simchowitz, 2013), 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 (Hoffman et al., 2013; Blei et al., 2016). 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, as compared with available methods. Since PCA is widely used as a primary step for further analysis, such as clustering, we also show how pCMF increases the performance of methods that are classically using PCA as a first step, especially the popular t-SNE (van der Maaten & Hinton, 2008; Amir et al., 2013). Using experimental published data, we show how pCMF provides a dimension reduction that is adapted to scRNA-seq data, by providing a better representation of the heterogeneities within datasets, which appears to be extremely helpful to characterize cell types. Finally, pCMF is available in the form of a R package available at `https://gitlab.inria.fr/gdurif/pCMF` (in beta version) and soon on the CRAN.

2 Results

We compare our method with standard approaches for unsupervised dimension reduction: the Poisson-NMF (Lee & Seung, 1999), applied to raw counts (model-based matrix factorization approach based on the Poisson distribution), and the sparse PCA (Witten et al., 2009) on log counts (based on an ℓ_1 penalty in the optimization problem defining the PCA to induce sparsity in the loadings $\mathbf{V}$). In addition, we use the Zero-Inflated Factor Analysis (ZIFA) by Pierson & Yau (2015), a dimension reduction approach that is specifically designed to handle dropout events in single-cell expression data (based on a zero-inflated Gaussian factor model applied to log-transformed counts). We present quantitative clustering results and qualitative visualization results on simulated and experimental scRNA-seq data. Another tool for dimension reduction and data visualization called t-SNE (van der Maaten & Hinton, 2008) is used for data visualization. It requires to choose a "perplexity" hyper-parameter that cannot be automatically calibrated, thus being less appropriate for a quantitative analysis.

2.1 Simulated data analysis

Details about data generation are given in appendix (c.f. Section A.3). We generate synthetic multivariate Negative Binomial counts, with $n = 100$ individuals and $p =$

1000 recorded variables. We artificially create clusters of individuals and groups of dependent variables. Then we set different levels of zero-inflation in the data (i.e. low or high probabilities of dropout events, corresponding to random null values in the data), and some part of the p variables are generated as random noise that do not induce any latent structure. Thus, we can test the performance of our method in different realistic data configurations.

2.1.1 Clustering in the observation space

Effect of zero-inflation. We first question the robustness of the different approaches to the level of zero-inflation (ZI) in the data (no ZI, low ZI, high ZI corresponding to a probability of dropout events being 0 or in $[0.4; 0.6]$ or in $[0.6; 0.8]$ respectively). We generate data with 3 groups of observations and train the different methods with $K = 3$ (fixed in this design). We also consider low and high separability between the groups of observations (c.f. Section A.3 in appendix). The quality of the clustering based on the reconstructed matrix $\widehat{\mathbf{U}}$ (see material and methods) will assess the ability of each method to retrieve the group structure in the observation space despite the dropout events. We measure the adjusted Rand Index (Rand, 1971) quantifying the accordance between the predicted clusters and original groups of individuals. Contrary to other approaches (Figure 1), pCMF adapts to the level of zero-inflation in the data and perfectly recovers the original groups of observations when the separability is high (adjusted Rand Index close to 1 in the different ZI configuration). The results of ZIFA indicates that using a zero-inflated Gaussian model is not sufficient to retrieve the groups in our count data. Indeed, methods based on transformed counts (like ZIFA and SPCA on log) do not account for the discrete nature of the data neither for their over-dispersion (O'Hara & Kotze, 2010). As for the Poisson-NMF method, its performance are comparable to pCMF for no dropout, but decrease as soon as there is zero-inflation in the data.

Effect of noisy genes. To quantify the impact of noisy genes on the retrieval of the individual groups, we consider data generated with different proportion (0%, 40% or 70%) of noisy genes that do not induce any structure in the data. We again consider two configurations where groups of individuals or lowly of highly separated (c.f. Section A.3 in appendix). The level of zero-inflation is set such that the probability of dropout events lies $[0.4; 0.6]$. In this setting, we train different models and compute the adjusted Rand Index for increasing values of K (number of components) to check the quality of the clustering of individuals when noisy genes are present and when introducing new components. Similarly to previous simulations, the clustering accuracy of pCMF is globally better than other methods, but all methods seem to be resilient to the addition of noisy genes, except for ZIFA whose performance decreases in this case. However, the performance of pCMF are not decreased by the introduction of new components, contrary to ZIFA or NMF, which means that our methods seems more robust to the choice of K.

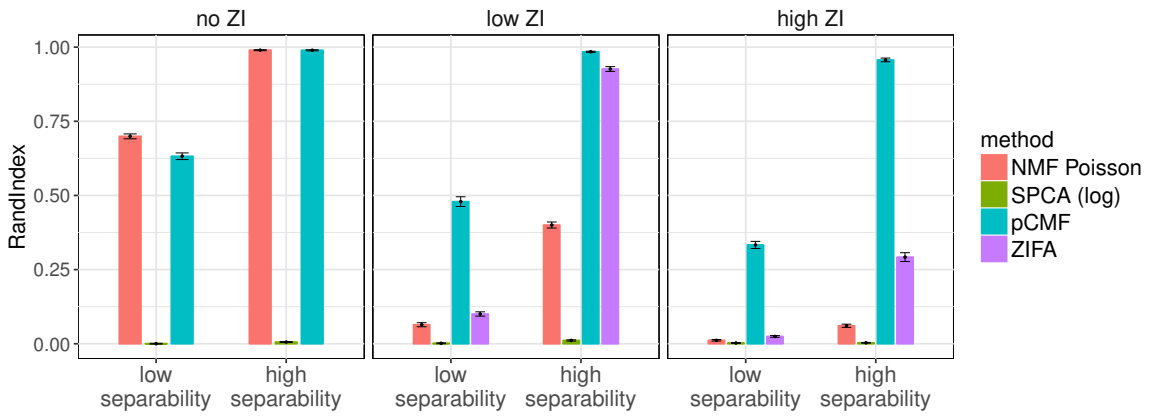

Figure 1: Adjusted Rand Index comparing clusters found by a κ-means algorithm (applied to $\widehat{\mathbf{U}}$ with $\kappa = 3$) and the original groups of individuals, for different levels of zero-inflation and different levels of separation between groups of individuals in the data. The number of components is set to $K = 3$. Data are generated with $n = 100$, $p = 1000$, 3 groups of individuals and 70% of noisy variables. Average values and deviation are estimated across 100 repetitions.

2.1.2 Data visualization

The question of the data visualization is central in many recent single-cell transcriptomic studies (e.g. Llorens-Bobadilla et al., 2015; Segerstolpe et al., 2016). The purpose is especially to represent a high dimensional data set in a low dimensional space that we can visualize (generally in 2 or 3 dimensions), in order to identify groups of cells or to illustrate the cell diversity. In the matrix factorization framework, we represent observation coordinates $(\widehat{u}_{i1}, \widehat{u}_{i2})_{i=1,\dots,n}$ from the matrix $\widehat{\mathbf{U}}$ when the dimension is $K = 2$ (see material and methods). We consider the same simulated data as previously ($n = 100$, $p = 1000$, 3 groups of observations, 70% of noisy variables, dropout probability in $[0.6; 0.8]$).

Our visual results are consistent with the previous clustering results (c.f. Figure 3). In this challenging context (high zero-inflation and numerous noisy variables), by using our pCMF approach, we are able to graphically identify the groups of individuals in the simulated zero-inflated count data. On the contrary, the 2-D visualization is not successful with the sparse PCA, ZIFA and Poisson-NMF, illustrating the interest of our data-specific approach compared to others. We mention that we represent the individual coordinates $\widehat{\mathbf{U}}$ in log scale for our method pCMF, because the natural representation associated to the Gamma distribution in the exponential family is the logarithm.

When considering t-SNE, it is generally used with a preliminary PCA step to reduce the dimension. It appears (c.f. Figure 3) that using our approach pCMF as a preliminary step before t-SNE gives better results for data visualization, This point supports our claim that using data-specific model improves the quality of the reconstruction in

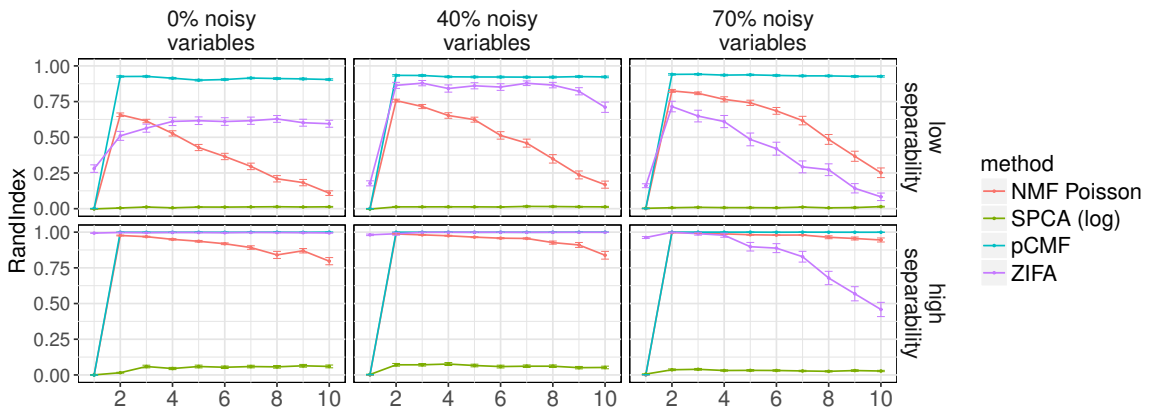

Figure 2: Adjusted Rand Index comparing clusters found by a κ-means algorithm (applied to $\widehat{\mathbf{U}}$ with $\kappa = 2$) and the original groups of individuals, depending on the number of components ($K = 1, \ldots, 10$), for different proportions of noisy genes and different levels of separation between groups of individuals in the data. Data are generated with $n = 100$, $p = 1000$, 2 groups of individuals and a probability of dropout events in $[0.4; 0.6]$. Average values and deviation are estimated across 100 repetitions.

the latent space. Here, we used $K = 20$ for the preliminary dimension reduction before t-SNE (both for PCA and pCMF). Using other dimensions (for instance $K = 50$ as in the default behavior of t-SNE) gives similar results.

2.1.3 Additional results

Additional results regarding computation time comparison with state-of-the-art approaches and performance enhancement of our zero-inflated sparse Gamma-Poisson factor model compared to standard Gamma-Poisson factor model are given in appendix (Section A.4). Although figures are not joined, we also mention that standard PCA does not give better quantitative or qualitative results than sparse PCA.

2.2 Analysis of single-cell data

We illustrate the performance of pCMF on a publicly available scRNA-seq dataset on neuronal stem cells (Llorens-Bobadilla et al., 2015). Neural stem cells (NCS) constitute an essential pool of adult cells for brain maintenance and repair. Llorens-Bobadilla et al. (2015) proposed a study to unravel the molecular heterogeneities of NCS populations based on scRNA-seq, and particularly focused on quiescent cells (qNSC). In their experiment, qNSC were transplanted in vivo in order to study their neurogenic activity. Following transplantation, 92 qNSC produced neuroblasts and olfactory neurons, whose transcriptome was compared with 21 astrocytes (CTX) and 27 transient amplifying progenitor cells (TAP). Then the authors used a PCA approach to reveal a continuum of "activation state", from astrocytes (low activation) to amplifying progenitor cells (TAP). We confront our pCMF output with the standard PCA, with ZIFA and with t-SNE results. The first visual result is that pCMF provides a better rep-

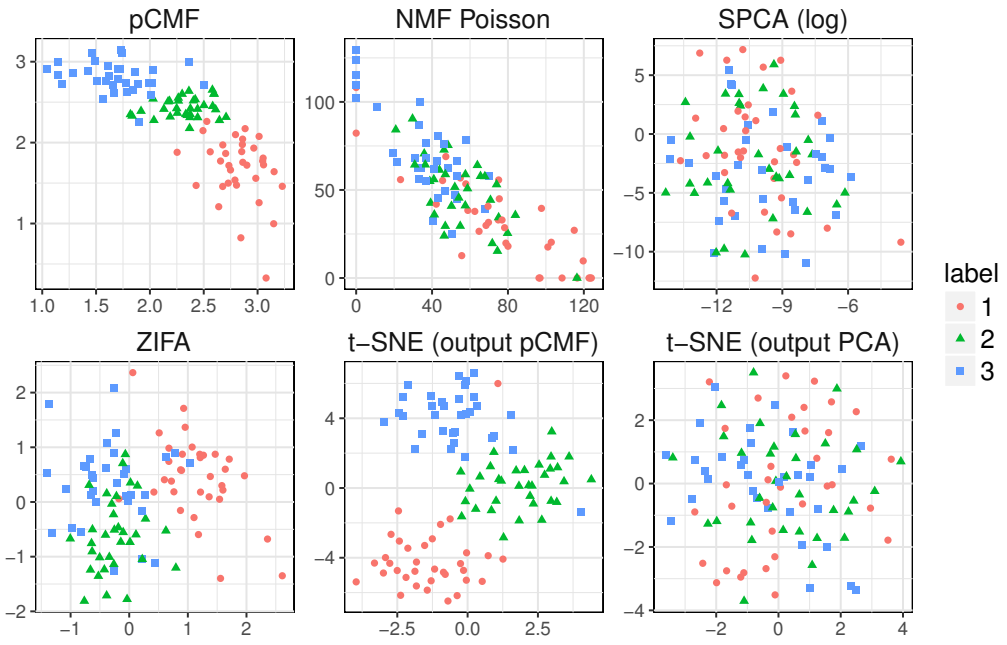

Figure 3: Individuals representation ($\widehat{\mathbf{U}}$) in a subspace of dimension $K = 2$. Data are generated with 3 groups of individuals, $n = 100$ and $p = 1000$, a probability of dropout events between 0.6 and 0.8, and 70% of noisy genes. t-SNE is applied with a preliminary dimension reduction step based on pCMF or PCA (default behavior) with $K = 20$.

resentation of the continuum than PCA and t-SNE, which probably reflects a better modeling of the biological variations that exist between activation states. Interestingly Llorens-Bobadilla et al. (2015) mention a minor overlap between qNSC and parenchymal astrocytes (CTX), whereas pCMF rather reveals an important overlap between CTX and qNSC1 cells. On the contrary, the t-SNE representation can hardly be interpreted as an activation continuum. The results from ZIFA are consistent with pCMF representation, which is a confirmation that the signal of this continuous activation state is strong in these data. Regarding the quantification of the biological variability, the first two axis of PCA only catches 11.74% of the total variance, whereas pCMF catches 69% of the total deviance. This illustrates the benefit of having a dimension reduction method that is based on the proper distribution and proper reduction quality metric.

3 Material and methods

We present the statistical model associated to our probabilistic Count Matrix Factorization (pCMF) approach, based on a zero-inflated sparse Gamma-Poisson factor model. Then, we introduce the framework to retrieve the factors $\mathbf{U}$ and $\mathbf{V}$ based on variational inference.

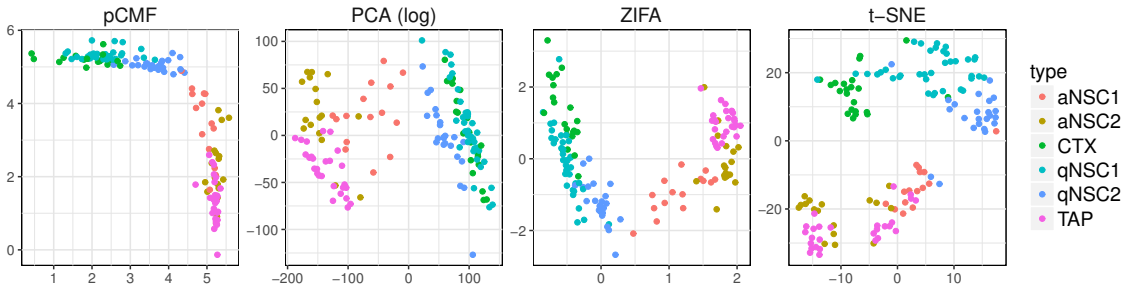

Figure 4: Analysis of the scRNA-seq data from Llorens-Bobadilla et al. (2015), $n = 141$ cells, $p \sim 14000$ genes. We pre-selected genes expressed (count > 1) in at least 2 cells, with a log-variance higher than 0.5 (as in the original paper). pCMF and t-SNE are applied to raw counts, while PCA and ZIFA are applied to log-transformed counts.

3.1 Count Matrix Factorization for zero-inflated over-dispersed data

Details about the model (construction, identifiability) are given in appendix (Section A.1).

Zero-Inflated Sparse Gamma-Poisson factor model. Our data consist in a matrix of counts, denoted by $\mathbf{X} \in \mathbb{N}^{n \times p}$, that we want to linearly decompose onto a subspace of dimension K, into a matrix product $\mathbf{U}\mathbf{V}^T$. The factor $\mathbf{U} \in \mathbb{R}^{+,n \times K}$ represent the coordinates of the observations (cells) in the subspace of dimension K, and $\mathbf{V} \in \mathbb{R}^{p \times K}$ the contributions (loadings) of variables (genes). In a standard Poisson Non-negative Matrix Factorization (NMF, Lee & Seung, 1999), the associated model verifies $\mathbf{X} \sim \mathscr{P}(\mathbf{U}\mathbf{V}^T)$. Details about the underlying geometry associated to the model (generalization of the Euclidean geometry with Bregman divergence and link with the deviance related to the model) are given in appendix (Section A.1.2).

To account for over-dispersion in the data, we consider the Gamma-Poisson representation (GaP, Cemgil, 2009; Zhou et al., 2012). To proceed, we consider a factor model, in which factors $\mathbf{U}$ and $\mathbf{V}$ are modeled as independent random variables with Gamma distributions such that $U_{ik} \sim \Gamma(\alpha_{k,1}, \alpha_{k,2})$ and $V_{jk} \sim \Gamma(\beta_{k,1}, \beta_{k,2})$.

To model zero-inflation (Simchowitz, 2013), we introduce a dropout indicator variable $D_{ij} \in \{0, 1\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. In this context, $D_{ij} = 0$ if gene j has been subject to a dropout event in cell i. Each D_{ij} follows a Bernoulli distribution with parameter π_j^{D}. The dropout indicators D_{ij} are assumed to be independent from the factors $\mathbf{U}$ and $\mathbf{V}$. Then, by integrating D_{ij} out, the conditional distribution of the counts is a zero-inflated Poisson distribution:

$$X_{ij} \,|\, \mathbf{U}_i, \mathbf{V}_j \sim \times(1 - \pi_j^{\mathrm{D}}) \times \delta_0 + \pi_j^{\mathrm{D}} \times \mathscr{P}\left(\textstyle\sum_k U_{ik} V_{jk}\right).$$

Finally we introduce some parsimony in our model, i.e. by assuming that only a proportion of recorded variables carry the signal, others being noise. To do so, the prior

on the loading variables V_{jk} is set to be a two-group sparsity-inducing prior (Engelhardt & Adams, 2014):

$$V_{jk} \sim (1 - \pi_j^{\mathrm{s}}) \times \delta_0 \; + \pi_j^{\mathrm{s}} \times \; \Gamma(\beta_{k,1}, \beta_{k,2}) \, ,$$

where π_j^{s} stands for the prior probability for gene j to contribute to any loading. This spike-and-slab formulation (Mitchell & Beauchamp, 1988) ensures that V_{jk} is either null (gene j does not contribute to factor k), or drawn from the Gamma distribution (when gene j contributes to the factor k).

Quality of the reconstruction. In our GaP model, we can use the deviance or equivalently the Bregman divergence (c.f. Section A.1.2 in appendix) between the data matrix $\mathbf{X}$ and the reconstructed matrix $\widehat{\mathbf{U}}\widehat{\mathbf{V}}^T$ to quantify the quality of the model. Regarding PCA, the percentage of explained variance is a natural and unequivocal quantification of the quality of the representation. In our case, since the models are not nested for increasing K, it appears non trivial to define a percentage of explained deviance.

We denote the conditional Poisson log-likelihood in our model as $\log p(\mathbf{X} \,|\, \boldsymbol{\lambda})$, where $\boldsymbol{\lambda}$ is a $n \times p$ matrix of Poisson intensities. To assess the quality of our model, we propose to define the percentage of explained deviance as:

$$\%\mathrm{dev} = \frac{\log p(\mathbf{X} \,|\, \boldsymbol{\lambda} = \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T) - \log p(\mathbf{X} \,|\, \boldsymbol{\lambda} = \bar{\mathbf{X}})}{\log p(\mathbf{X} \,|\, \boldsymbol{\lambda} = \mathbf{X}) - \log p(\mathbf{X} \,|\, \boldsymbol{\lambda} = \bar{\mathbf{X}})}$$

where $\widehat{\mathbf{U}}\widehat{\mathbf{V}}^T$ is the predicted reconstructed matrix in our model, and $\bar{\mathbf{X}}$ is the column average of $\mathbf{X}$. We use two baselines: (i) the log-likelihood of the saturated model, i.e. $\log p(\mathbf{X} \,|\, \boldsymbol{\lambda} = \mathbf{X})$ (as in the deviance), which corresponds to the richest model and (ii) the log-likelihood of the model where each Poisson intensities λ_{ij} is estimated by the average of the observations in the column j, i.e. $\log p(\mathbf{X} \,|\, \boldsymbol{\lambda} = \bar{\mathbf{X}})$, which is the most simple model that we could use. This formulation ensures that the ratio %dev lies in $[0; 1]$.

3.2 Factor inference.

To avoid using the heavy machinery of MCMC (Nathoo et al., 2013) to infer the intractable posterior of the latent variables in our model, we use the framework of variational inference (Hoffman et al., 2013). The principle is to approximate the intractable posterior by a factorizable distribution, called the variational distribution, regarding the Kullback-Leibler divergence (that quantify probability distribution proximity). Variational inference can be reformulated into a maximization problem, that admits a solution under some reasonable assumptions on the variational distributions.

To be more precise, we use a variational EM algorithm (Beal & Ghahramani, 2003) that allows to jointly approximate the posterior distributions of the latent variables and the hyper-parameters of the model. This approach was successfully adapted to

the standard Gamma-Poisson factor model Dikmen & Févotte (2012), and we propose an extension to our zero-inflated sparse model. Details about the inference framework are given in appendix (Section A.2).

4 Conclusion

In this work, we provide a new framework for dimension reduction in unsupervised context. In particular, we introduce a model-based matrix factorization method specifically designed to analyse single-cell RNA-seq data. Our probabilistic Count Matrix Factorization (pCMF) approach accounts for the specificity of these data, being zero-inflated and over-dispersed counts. In other word, we propose a generalized PCA procedure that is suitable for data visualization and clustering. The interest of our zero-inflated sparse Gamma-Poisson factor model is to replace the variance-based formulation of PCA, associated to the Euclidean geometry and the Gaussian distribution, with a metric (based on Bregman divergence) that is adapted to scRNA-seq data characteristics.

Analyzing single-cell expression profiles is a huge challenge to understand the cell diversity in a tissue/an organism and more precisely characterize the associated gene activity. We show on simulations and experimental data that our pCMF approach is able to catch the underlying structure in zero-inflated over-dispersed count data. In particular, we show that our method can be used for data visualization in a lower dimensional space or for preliminary dimension reduction before a clustering step. In both cases, pCMF performs as well or out-performs state-of-the-art approaches, especially the PCA (being the gold standard) or more specific methods such as the NMF (count based) or ZIFA (zero-inflation specific).

In addition, our work could benefit from improvements. We are working on a model selection strategy to automatically select the dimension K, based on the integrated completed likelihood (Matthieu & Mohammed, 2016). This could refine the use of pCMF as a preliminary dimension reduction step before clustering or visualization with t-SNE.

Funding

This work was supported by the french National Resarch Agency (ANR) as part of the "Algorithmics, Bioinformatics and Statistics for Next Generation Sequencing data analysis" (ABS4NGS) ANR project [grant number ANR-11-BINF-0001-06] and as part of the "MACARON" ANR project [grant number ANR-14-CE23-0003]. It was performed using the computing facilities of the computing center LBBE/PRABI.

References

Amir, e. l. A. D., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., & Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat. Biotechnol.*, *31*(6), 545–552.

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with Bregman Divergences. *Journal of Machine Learning Research*, *6*(Oct), 1705–1749.

Beal, M. J., & Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. *Bayesian statistics*, *7*, 453–464.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2016). Variational Inference: A Review for Statisticians. *arXiv preprint arXiv:1601.00670*.

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., & Greenleaf, W. J. (2015). Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, *523*(7561), 486–490.

Buettner, F., Natarajan, K. N., Casale, F. P., Proserpio, V., Scialdone, A., Theis, F. J., Teichmann, S. A., Marioni, J. C., & Stegle, O. (2015). Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. *Nat. Biotechnol.*, *33*(2), 155–160.

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

Chen, P., Chen, Y., & Rao, M. (2008). Metrics defined by Bregman Divergences. *Communications in Mathematical Sciences*, *6*(4), 915–926.

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).

Dikmen, O., & Févotte, C. (2012). Maximum marginal likelihood estimation for nonnegative dictionary learning in the Gamma-Poisson model. *Signal Processing, IEEE Transactions on*, *60*(10), 5163–5175.

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218.

Engelhardt, B. E., & Adams, R. P. (2014). Bayesian Structured Sparsity from Gaussian Fields. *arXiv:1407.2235 [q-bio, stat]*.

Févotte, C., & Cemgil, A. T. (2009). Nonnegative matrix factorizations as probabilistic inference in composite models. In *Signal Processing Conference, 2009 17th European*, (pp. 1913–1917). IEEE.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., & Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531–537.

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic Variational Inference. *J. Mach. Learn. Res.*, *14*(1), 1303–1347.

Landgraf, A. J., & Lee, Y. (2015). Generalized principal component analysis: Projection of saturated model parameters. *Technical Report 892, Department of Statistics, The Ohio State University*.

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

Llorens-Bobadilla, E., Zhao, S., Baser, A., Saiz-Castro, G., Zwadlo, K., & Martin-Villalba, A. (2015). Single-Cell Transcriptomics Reveals a Population of Dormant Neural Stem Cells that Become Activated upon Brain Injury. *Cell Stem Cell*, *17*(3), 329–340.

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

Matthieu, M., & Mohammed, S. (2016). Variable selection for model-based clustering using the integrated complete-data likelihood. *Statistics and Computing*.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Nathoo, F. S., Lesperance, M. L., Lawson, A. B., & Dean, C. B. (2013). Comparing variational Bayes with Markov chain Monte Carlo for Bayesian computation in neuroimaging. *Statistical methods in medical research*, *22*(4), 398–423.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122.

Paul, F., Arkin, Y., Giladi, A., Jaitin, D. A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F. K., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B. T., Tanay, A., & Amit, I. (2015). Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell*, *163*(7), 1663–1677.

Picelli, S., Bjorklund, A. K., Faridani, O. R., Sagasser, S., Winberg, G., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods*, *10*(11), 1096–1098.

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

Rand, W. M. (1971). Objective Criteria for the Evaluation of Clustering Methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Rotem, A., Ram, O., Shoresh, N., Sperling, R. A., Goren, A., Weitz, D. A., & Bernstein, B. E. (2015). Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.*, *33*(11), 1165–1172.

Segerstolpe, Å., Palasantza, A., Eliasson, P., Andersson, E.-M., Andréasson, A.-C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M. K., Smith, D. M., Kasper, M., Ämmälä, C., & Sandberg, R. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metabolism*, *24*(4), 593–607.

Simchowitz, M. (2013). Zero-Inflated Poisson Factorization for Recommendation Systems. *Junior Independent Work (advised by D. Blei), Princeton University, Department of Mathematics*.

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).

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., & Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.*, *32*(4), 381–386.

van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Wagner, A., Regev, A., & Yosef, N. (2016). Revealing the vectors of cellular identity with single-cell genomics. *Nat. Biotechnol.*, *34*(11), 1145–1160.

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, *10*(3), 515–534.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, *151*(1), 3–34.

Zhou, M., & Carin, L. (2012). Augment-and-Conquer Negative Binomial Processes. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.) *Advances in Neural Information Processing Systems 25*, (pp. 2546–2554). Curran Associates, Inc.

Zhou, M., Hannah, L. A., Dunson, D. B., & Carin, L. (2012). Beta-negative binomial process and Poisson factor analysis. In *In AISTATS*.

Zong, C., Lu, S., Chapman, A. R., & Xie, X. S. (2012). Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, *338*(6114), 1622–1626.

Appendix

A.1 Count Matrix Factorization for zero-inflated over-dispersed data

A.1.1 The Poisson factor model

Our data consist in a matrix of counts, denoted by $\mathbf{X} \in \mathbb{N}^{n \times p}$, that we want to decompose onto K principal components, K being fixed in a first step. We introduce $\mathbf{U} \in \mathbb{R}^{+,n \times K}$ the coordinates of the observations (cells) on the K principal components, and $\mathbf{V} \in \mathbb{R}^{p \times K}$ the contributions (loadings) of variables (genes) on the new axes. In a standard Poisson Matrix Factorization (see Lee & Seung, 1999), that we call Poisson-NMF, the model is such that $\mathbf{X} \sim \mathscr{P}\big(\mathbf{U}\mathbf{V}^T\big)$.

A.1.2 Underlying geometry

To quantify the quality of approximation of matrix $\mathbf{X}$ by $\mathbf{\Lambda} = \mathbf{U}\mathbf{V}^T$, we consider the Bregman divergence as a metric (see Banerjee et al., 2005; Chen et al., 2008). This divergence can be viewed as a generalization of the Euclidean metric to the exponential family. Thus the model we propose is developed within the framework of the generalized PCA proposed by Collins et al. (2001) and based on this Bregman divergence. In the Poisson model, the Bregman divergence between two $n \times p$ matrices $\mathbf{X}$ and $\mathbf{\Lambda}$ is defined as (Févotte & Cemgil, 2009):

$$D(\mathbf{X} \,|\, \mathbf{\Lambda}) = \sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} \, \log\left(\frac{x_{ij}}{\Lambda_{ij}}\right) - x_{ij} + \Lambda_{ij}.$$

The interest here is to choose a geometry that is induced by an appropriate probabilistic model dedicated to count data. Indeed, the least squares criterion used in PCA for instance, might not be appropriate for non-Gaussian data. The Bregman divergence can also be related to the deviance of the Poisson model defined such as

$$\text{Dev}(\mathbf{X}, \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T) = -2 \times \left(\log p(\mathbf{X} \,|\, \mathbf{\Lambda} = \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T) - \log p(\mathbf{X} \,|\, \mathbf{\Lambda} = \mathbf{X}) \right),$$

with $\log p(\mathbf{X} \,|\, \mathbf{\Lambda})$ the Poisson log-likelihood, thus $\text{Dev}(\mathbf{X}, \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T) \propto D(\mathbf{X} \,|\, \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T)$.

A.1.3 Modeling over-dispersion

To account for over-dispersion in the data, we consider the Poisson Gamma representation (GaP), as proposed by Cemgil (2009). To proceed, we consider a factor model, in which factors $\mathbf{U}$ and $\mathbf{V}$ are modeled as independent random variables with Gamma distributions such that

$$\begin{aligned} U_{ik} &\sim \Gamma(\alpha_{k,1}, \alpha_{k,2}), \\ V_{jk} &\sim \Gamma(\beta_{k,1}, \beta_{k,2}). \end{aligned} \tag{A.1}$$

Then some third-party latent variables are introduced to facilitate the derivation of our inference methods. We consider latent variables $\mathbf{Z} = [Z_{ijk}] \in \mathbb{R}^{n \times p \times K}$, defined such that $X_{ij} = \sum_k Z_{ijk}$. This new indicator variable quantifies the contribution of factor k to the data. Here Z_{ijk} are assumed to be conditionally independent and to follow a conditional Poisson distribution, i.e. $Z_{ijk} \,|\, U_{ik}, V_{jk} \sim \mathscr{P}(U_{ik}\, V_{jk})$. Thus, the conditional distribution of X_{ij} remains $\mathscr{P}(\sum_k U_{ik}\, V_{jk})$ thanks to the additive property of the Poisson distribution.

A.1.4 Dropout modeling using a zero-inflated (ZI) model

We introduce a dropout variable $D_{ij} \in \{0, 1\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$. This indicator is defined such that each $D_{ij} = 0$ if gene j has been subject to a dropout event in cell i, with $D_{ij} \sim \mathcal{B}(\pi_j^{\mathrm{D}})$. We consider gene-specific dropout rates, π_j^{D}, following recommendations of the literature (Pierson & Yau, 2015). Thus, to include zero inflation in the probabilistic factor model, we consider that:

$$X_{ij} \,|\, \mathbf{U}_i, \mathbf{V}_j, \mathbf{D} \sim \times(1 - D_{ij}) \times \delta_0 + D_{ij} \times \mathscr{P}\big(\sum_k U_{ik}\, V_{jk}\big)\,.$$

The dropout indicators D_{ij} are assumed to be independent from the factors. Then we can check, by integrating D_{ij} out, that the probability of observing a zero in the data becomes:

$$\mathbb{P}\big(X_{ij} = 0 \,|\, \mathbf{U}_i, \mathbf{V}_j\,;\, \boldsymbol{\pi}\big) = (1 - \pi_j^{\mathrm{D}}) + \pi_j^{\mathrm{D}}\, \exp\big(-\textstyle\sum_k U_{ik}\, V_{jk}\big)\,,$$

which illustrates the two potential sources of zeros.

A.1.5 Probabilistic variable selection

Finally we suppose that our model is parsimonious, by considering that among all recorded variables, only a proportion carries the signal, the others being noise. To do so, we modify the prior of the loadings variables V_{jk}, to consider a sparse model with a two-group sparsity-inducing prior. The model is then enriched by the introduction of a new indicator variable $S_{jk} \sim \mathcal{B}(\pi_j^{\mathrm{S}})$, that equals 1 if gene j contributes to the loading V_{jk}, and zero otherwise. π_j^{S} stands for the prior probability for gene j to contribute to any loading. To define the sparse GaP factor model, we modify the distribution of the loadings latent factor V_{jk}, such that

$$V_{jk} | S_{jk} \sim (1 - S_{jk}) \times \delta_0 \; + S_{jk} \times \; \Gamma(\beta_{k,1}, \beta_{k,2})\,.$$

This spike-and-slab formulation ensures that V_{jk} is either null (gene j does not contribute to factor k), or drawn from the Gamma distribution (when gene j contributes to the factor). Then the contribution of gene j to the component k is accounted for in the conditional Poisson distribution of X_{ij}, with

$$X_{ij} \,|\mathbf{U}_i, \mathbf{V}'_j, \mathbf{D}, \mathbf{S}_j \sim (1 - D_{ij})(1 - S_{jk}) \times \delta_0 + \mathscr{P}\big(D_{ij} \textstyle\sum_k U_{ik}\, [S_{jk}\, V'_{jk}]\big),$$

where $V_{jk} = S_{jk}\, V'_{jk}$ such that $V'_{jk} \sim \Gamma(\beta_{k,1}, \beta_{k,2})$.

A.1.6 Model Identifiability

Scaling effect. As stated in Dikmen & Févotte (2012), GaP factor models suffer from identifiability issues, due to the scaling of the Gamma prior parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Indeed, considering $\alpha^*_{k,2} = \eta_k\,\alpha_{k,2}$ and $\beta^*_{k,2} = \eta_k^{-1}\,\beta_{k,2}$ for fixed values η_k, and usins the scaling property of the Gamma distribution: if $U_{ik} \sim \text{Gamma}(\alpha_{k,1}, \alpha_{k,2})$ then $\eta_k U \sim \Gamma(\alpha_{k,1}, \eta_k^{-1}\alpha_{k,2})$. We can show that the joint log-likelihood regarding $\mathbf{U}\mathbf{H}^{-1}$ and $\mathbf{V}\mathbf{H}$ with $\mathbf{H} = \text{diag}(\eta_k)_{k=1:K}$ verifies:

$$\begin{aligned} &\log p(\mathbf{X}, \mathbf{U}\mathbf{H}^{-1}, \mathbf{V}\mathbf{H} \,|\, \boldsymbol{\alpha}_1, \mathbf{H}\boldsymbol{\alpha}_2, \boldsymbol{\beta}_1, \mathbf{H}^{-1}\boldsymbol{\beta}_2) \\ &= \quad \log p(\mathbf{X}, \mathbf{U}, \mathbf{V} \,|\, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \\ &\qquad + (n-p)\sum_k \log(\eta_k) \end{aligned} \tag{A.2}$$

When $n = p$, there is an identifiability issue regarding the scaling of the parameters $\alpha_{k,2}$ and $\beta_{k,2}$, because different values lead to the same joint log-likelihood. In such case, a solution will be to fix the scale parameters $\alpha_{k,2}$ and $\beta_{k,2}$ to avoid the scaling effect. When $n \neq p$, the only problem is a potential solution with infinite norm with $\alpha_{k,2} \to 0$ and $\beta_{k,2} \to \infty$ or vice-versa (c.f. Dikmen & Févotte, 2012). However, in practice we did not encounter such sequence of diverging parameters.

When considering sparsity and/or zero-inflation in the model, Equation (A.2) still holds regarding the parameters of the Gamma prior distributions and we have to consider the same precaution.

Factor order. In practice, principal components of standard PCA show very convenient properties: they are orthogonal (thanks to the SVD), they can be naturally ordered (thanks to Eckart & Young (1936) theorem), and they are associated to nested models. Unfortunately, likelihood-based factor models do not share the same properties. Indeed, for NMF or for our GaP factor model, the likelihood that defines the model is identifiable up to a permutation of factors (i.e. by permuting the columns in $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ according to the same reordering). Hence, there does not exist a natural ordering for components of probabilistic factor models. Thus we propose an ordering defined by the cumulative Bregman divergence:

$$k \mapsto D\big(\mathbf{X} \,|\, \widehat{\mathbf{U}}_{1:k}(\widehat{\mathbf{V}}_{1:k})^T\big)\,.$$

In addition, we mention that the different GaP factor models are not nested when the dimension K increases (as in the NMF), thus the factor estimates are computed for any dimension, contrary to PCA.

A.2 Model inference using a variational EM algorithm

Our goal is to infer the posterior distributions over the factors $\mathbf{U}$ and $\mathbf{V}$ depending on the data $\mathbf{X}$. To proceed, we extend the version of the variational EM algorithm (Beal & Ghahramani, 2003) proposed by Dikmen & Févotte (2012) in the context of the standard Gamma-Poisson factor model, to our sparse and zero-inflated GaP model.

A.2.1 Definition of variational distributions

In the variational framework, the posterior $p(\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{D} \,|\, \mathbf{X})$ is approximated by the variational distribution $q(\mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{D})$ regarding the Kullback-Leibler divergence (Hoffman et al., 2013), that quantifies the divergence between two probability distributions. Since the posterior is not explicit, the inference of q is based on the optimization of the Evidence Lower Bound (ELBO), denoted by $J(q)$ and defined as:

$$J(q) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z}, \mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{D})] - \mathbb{E}_q[\log q(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \mathbf{S}, \mathbf{D})]\,, \tag{A.3}$$

that is a lower bound on the marginal log-likelihood $\log p(\mathbf{X})$. In addition, maximizing the ELBO $J(q)$ is equivalent to minimizing the KL divergence between q and the posterior distribution of the model (Hoffman et al., 2013).

To derive the optimization, q is assumed to lie in the mean-field variational family, i.e. (i) to be factorisable with independence between latent variables and between observations and (ii) to follow the conjugacy in the exponential family, i.e. to be in the same exponential family as the full conditional distribution on each latent variables in the model.

Thanks to the first assumption, in our model, the variational distribution q is defined as follows:

$$\begin{aligned} q(\mathbf{U}, \mathbf{V}, \mathbf{Z}, \mathbf{S}, \mathbf{D}) = \; & \prod_{i=1}^{n}\prod_{k=1}^{K} q(U_{ik} \,|\, \mathbf{a}_{ik}) \times \prod_{j=1}^{p}\prod_{k=1}^{K} q(V'_{jk} \,|\, \mathbf{b}_{jk}) \\ & \times \prod_{i=1}^{n}\prod_{j=1}^{p} q\big((Z_{ijk})_k \,|\, (R_{ijk})_k\big) \times \prod_{j=1}^{p}\prod_{k=1}^{K} q(S_{jk} \,|\, p^{\mathrm{S}}_{jk}) \\ & \times \prod_{i=1}^{n}\prod_{j=1}^{p} q(D_{ij} \,|\, p^{\mathrm{D}}_{ij}) \end{aligned}$$

where $\mathbf{a}_{ik}$, $\mathbf{b}_{jk}$, $(r_{ijk})_k$, p^{S}_{jk} and p^{D}_{ij} are the parameters of the variational distribution regarding U_{ik}, V'_{jk}, $(Z_{ijk})_k$, S_{jk}, D_{ij}, respectively. Then we need to precise the full conditional distributions of the model before defining the variational distributions more precisely.

A.2.2 Full conditional distributions

In our factor model all full conditionals are tractable. Thanks to the Gamma-Poisson conjugacy, the full conditionals of U_{ik} and V'_{jk} are Gamma distributions. The proof is based on the Bayes rule and the distribution of the latent variables $\mathbf{Z}$, that are actually necessary to derive $p(U_{ik} \,|\, —)$ and $p(V'_{jk} \,|\, —)$. The full conditional of the vector $\mathbf{Z}_{ij}$ is also explicit, being a Multinomial distribution (Zhou & Carin, 2012) when $D_{ij} \neq 0$ and deterministic null when $D_{ij} = 0$, i.e. $(Z_{ijk})_k \,|\, — \;\sim D_{ij}\, \mathcal{M}\big(X_{ij}, (\rho_{ijk})_k\big)$. Here

the Multinomial probabilities $(\rho_{ijk})_k$ depend on $(S_{jk}, U_{ik}, V'_{jk})_k$, and quantify the prior contribution of factor k to the observations X_{ij}, i.e.

$$\rho_{ijk} = \frac{S_{jk}\, U_{ik}\, V'_{jk}}{\sum_\ell S_{j\ell}\, U_{i\ell}\, V'_{j\ell}}.$$

This point justifies why the variational distribution is based on the vector $\mathbf{Z}_{ij}$ instead of taking each Z_{ijk} separately. Note that if the S_{jk} are null for all k or if $D_{ij} = 0$ (i.e. $X_{ij} = 0$), the vector $(Z_{ijk})_k$ is deterministic and takes null values.

We summarize the full conditionals in the sparse ZI-GaP factor model regarding U_{ik}, V_{jk} and $(Z_{ijk})_k$, that are defined such as:

$$\begin{aligned} U_{ik}\,|\,— &\sim \Gamma(\alpha_{k,1} + \textstyle\sum_j D_{ij}\, S_{jk}\, Z_{ijk},\ \alpha_{k,2} + \sum_j D_{ij}\, S_{jk}\, V_{jk})\,, \\ V_{jk}\,|\,— &\sim \Gamma(\beta_{k,1} + \textstyle\sum_i D_{ij}\, S_{jk}\, Z_{ijk},\ \beta_{k,2} + \sum_i D_{ij}\, S_{jk}\, U_{ik})\,, \\ (Z_{ijk})_k\,|\,— &\sim D_{ij}\, \mathcal{M}\Big(X_{ij}, (\rho_{ijk})_k\Big)\,, \end{aligned} \tag{A.4}$$

Zero Inflation. Regarding the zero-inflation indicators, D_{ij} is a binary variable, its distribution is either deterministic or Bernoulli. When the entry X_{ij} is non null, D_{ij} is certainly equal to one. When $X_{ij} = 0$, the full conditional is explicit and the Bernoulli probability only depends on the prior over D_{ij} and the probability that X_{ij} is null. It can be formulated as follows:

$$p(D_{ij} = 1\,|\,—) = \frac{\pi_j^{\mathrm{D}}\ e^{-\sum_k S_{jk}\, U_{ik}\, V'_{jk}}}{(1 - \pi_j^{\mathrm{D}}) + \pi_j^{\mathrm{D}}\, e^{-\sum_k S_{jk}\, U_{ik}\, V'_{jk}}}\,.$$

Sparsity and variable selection. The sparsity indicator S_{jk} is also a binary variable and its full conditional is also an explicit Bernoulli distribution. It depends on the prior over S_{jk} and the probability that gene j contributes to the components k, quantified by the joint distribution on $(Z_{ijk})_i$, thus:

$$p(S_{jk} = 1\,|\,—) \propto\ \ \pi_j^{\mathrm{s}} \times \textstyle\prod_i \exp(-S_{jk}\, U_{ik}\, V'_{jk})\ (S_{jk}\, U_{ik}\, V'_{jk})^{Z_{ijk}}.$$

A.2.3 Approximate posteriors

To approximate the (intractable) posterior distributions, variational distributions are assumed to lie in the same exponential family as the corresponding full conditionals and to be independent such that:

$$\mathbf{Z}_{ij} \overset{q}{\sim} \mathcal{M}\Big((r_{ijk})_k\Big) \qquad \begin{aligned} U_{ik} &\overset{q}{\sim} \Gamma(a_{ik,1}, a_{ik,2}) & S_{jk} &\overset{q}{\sim} \mathcal{B}(p_{jk}^{\mathrm{s}}) \\ V'_{jk} &\overset{q}{\sim} \Gamma(b_{jk,1}, b_{jk,2}) & D_{ij} &\overset{q}{\sim} \mathcal{B}(p_{ij}^{\mathrm{D}}), \end{aligned}$$

where $\overset{q}{\sim}$ denotes the variational distribution. The strength of our approach is the resulting explicit approximate distribution on the loadings that induces sparsity:

$$V_{jk}|S_{jk} \overset{q}{\sim} (1 - S_{jk}) \times \delta_0 + S_{jk} \times \Gamma(b_{jk,1}, b_{jk,2}),$$

In the following, the derivation of variational parameters involves the moments and log-moments of the latent variables regarding the variational distribution. Since the distributions q is fully determined, these moments can be directly computed. For the sake of simplicity, we will use notation $\widehat{U}_{ik} = \mathbb{E}_q[U_{ik}]$ and $\widehat{\log U}_{ik} = \mathbb{E}_q[\log U_{ik}]$ (collected in the matrices $\widehat{\mathbf{U}}$ and $\widehat{\log \mathbf{U}}$ respectively), with similar notations for other hidden variables of the model (V_{jk}, D_{ij}, S_{jk}, Z_{ijk}).

A.2.4 Derivation of variational parameters

In order to find a stationary point of the ELBO, $J(q)$ is differentiated regarding each variational parameter separately. The formulation of the ELBO regarding each parameter separately is based on the corresponding full conditional, i.e. $p(U_{ik} \,|\, —)$, $p(V_{jk} \,|\, —)$ and $p\big((Z_{ijk})_k \,|\, —\big)$. The partial formulation are therefore respectively:

$$
\begin{aligned}
J(q)\big|_{\mathbf{a}_{ik}} &= \mathbb{E}_q[\log p(U_{ik} \,|\, —)] - \mathbb{E}_q[\log q(U_{ik}\,;\, \mathbf{a}_{ik})] + \text{cst} \\
J(q)\big|_{\mathbf{b}_{jk}} &= \mathbb{E}_q[\log p(V_{jk} \,|\, —)] - \mathbb{E}_q[\log q(V_{jk}\,;\, \mathbf{b}_{jk})] + \text{cst} \\
J(q)\big|_{(r_{ijk})_k} &= \mathbb{E}_q\big[\log p\big((Z_{ijk})_k \,|\, —\big)\big] - \mathbb{E}_q\big[\log q\big((Z_{ijk})_k\,;\, (r_{ijk})_k\big)\big] + \text{cst}
\end{aligned}
$$

Therefore, the ELBO is explicit regarding each variational parameter and the gradient of the ELBO $J(q)$ depending on the variational parameters $\mathbf{a}_{ik}$, $\mathbf{b}_{jk}$ and r_{ijk} respectively can be derived to find the coordinate of the stationary point, that corresponds to a local optimum. In practice, the optimum value for each variational parameter corresponds to the expectation regarding q of the corresponding parameter of the full conditional distribution (Hoffman et al., 2013). Thus the coordinates of the ELBO's gradient optimal point are explicit.

Variational parameters of factors. We derive the stationary point formulation for the variational parameters regarding U_{ik} and V_{jk}, being explicitly (directly derived from the partial derivatives of $J(q)$):

$$
\begin{aligned}
\mathbf{a}_{ik} &= \Big(\alpha_{k,1} + \textstyle\sum_j \widehat{D}_{ij}\, \widehat{S}_{jk}\, \widehat{Z}_{ijk}\,,\ \alpha_{k,2} + \sum_j \widehat{D}_{ij}\, \widehat{S}_{jk}\, \widehat{V'}_{jk}\Big)^T \\
\mathbf{b}_{jk} &= \Big(\beta_{k,1} + \widehat{S}_{jk} \textstyle\sum_i \widehat{D}_{ij}\, \widehat{Z}_{ijk}\,,\ \beta_{k,2} + \widehat{S}_{jk} \sum_i \widehat{D}_{ij}\, \widehat{U}_{ik}\Big)^T.
\end{aligned}
$$

As for variable $Z_{ijk} = U_{ij}V_{jk}$, its posterior distribution depends on parameter r_{ijk} with the relation $\log(r_{ijk}) = \mathbb{E}_q[\log(\rho_{ijk})]$. Hence, the variational distribution on $(Z_{ijk})_k$ naturally depends on the selection indicator S_{jk} (since our model focuses on loadings selection). In particular, the variational parameter r_{ijk} depends on S_{jk}, trough a specific term $\mathbb{E}_q[\log(S_{jk}\, V'_{jk})]$ that is computed using the variational distribution of S_{jk} (a Bernoulli distribution of parameter p^{s}_{jk}). To proceed, we introduce $\widetilde{S}_{jk}$, the discretized predictor of S_{jk} such that

$$
\widetilde{S}_{jk} = \mathbf{1}_{\{p^{\text{s}}_{jk} > \tau\}},
$$

where τ is a threshold specified by the user (for instance 0.5). Then, the formulation of the optimal variational parameter r_{ijk} is approximated by:

$$r_{ijk} = \frac{\widetilde{S}_{jk} \exp\left(\widehat{\log U}_{ik} + \widehat{\log V'}_{jk}\right)}{\sum_\ell \widetilde{S}_{j\ell} \exp\left(\widehat{\log U}_{i\ell} + \widehat{\log V'}_{j\ell}\right)}.$$

Variational dropout proportion. Regarding the zero-inflated probabilities p^{D}_{ij}, when $X_{ij} \neq 0$, the posterior is explicit since $D_{ij} = 1$ with probability one. Hence, only the case $X_{ij} = 0$ requires a variational inference. As stated previously, the full conditional is explicit and it is possible to derive and optimize the ELBO (based on the natural parametrization of the Bernoulli distribution in the exponential family). Eventually, p^{D}_{ij} is computed as:

$$\mathrm{logit}(p^{\mathrm{D}}_{ij}) = \ \mathrm{logit}(\pi^{\mathrm{D}}_j) - \textstyle\sum_k \widehat{S}_{jk}\, \widehat{U}_{ik} \widehat{V}'_{jk}\,,$$

where the Bernoulli prior probability π^{D}_j is corrected by $\mathbb{E}_q[\log \mathbb{P}(X_{ij} = 0)]$ to account for the probability of X_{ij} being a true zero.

Variational Selection probability. Concerning the sparse indicator S_{jk}, the natural parametrization of the Bernoulli distribution is based on the logit of the Bernoulli probability. Hence we can write an explicit formulation of the ELBO regarding p^{S}_{jk} based on the full conditional on S_{jk}. Following this formulation, the stationary point p^{S}_{jk} verifies:

$$\mathrm{logit}(p^{\mathrm{S}}_{jk}) = \mathrm{logit}(\pi^{\mathrm{S}}_j) - \textstyle\sum_i \widehat{D}_{ij}\, \widehat{U}_{ik}\, \widehat{V}'_{jk} + \widehat{D}_{ij}\, \widehat{Z}_{ijk} \left(\widehat{\log U}_{ik} + \widehat{\log V'}_{jk}\right).$$

This corresponds to a correction of the Bernoulli prior probability π^{S}_j, depending is on the quantification of the contribution of gene j to component k in all individuals, i.e. $\mathbb{E}_q[\sum_i \log p(Z_{ijk})]$.

A.2.5 Variational EM algorithm

We use the variational-EM algorithm (Beal & Ghahramani, 2003) to jointly approximate the posterior distributions and to estimate the hyper-parameters $\boldsymbol{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}^{\mathrm{S}}, \boldsymbol{\pi}^{\mathrm{D}})$. In this framework, the variational inference is used within a variational E-step, in which the standard expectation of the joint likelihood regarding the posterior $\mathbb{E}[p(\mathbf{X}, \mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{D}\,; \boldsymbol{\Omega})|\mathbf{X}]$ is approximated by

$$\mathbb{E}_q[p(\mathbf{X}, \mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{D}\,; \boldsymbol{\Omega})].$$

Then the variational M-step consists in maximizing $\mathbb{E}_q[p(\mathbf{X}, \mathbf{U}, \mathbf{V}, \mathbf{S}, \mathbf{D}\,; \boldsymbol{\Omega})]$ w.r.t. the hyper-parameters $\boldsymbol{\Omega}$. In the variational-EM algorithm, we have explicit formulations of the stationary points regarding variational parameters (E-step) and prior hyper-parameters (M-step) in the model, thus we use a coordinate descent iterative algorithm (see Wright, 2015, for a review) to infer the variational distribution.

In particular, the hyper-parameters are updated within the M-step such that:

$$\begin{aligned}
\alpha_{k,1} &= \psi^{-1}\left(\log\alpha_{k,2} + \frac{1}{n}\sum_i \widehat{\log U}_{ij}\right), \quad \alpha_{k,2} = \frac{\alpha_{k,1}}{\sum_i \widehat{U}_{ij}/n}, \\
\beta_{k,1} &= \psi^{-1}\left(\log\beta_{k,2} + \frac{1}{p}\sum_j \widehat{\log V}_{ij}\right), \quad \beta_{k,2} = \frac{\beta_{k,1}}{\sum_j \widehat{V}_{ij}/p}, \\
\pi_j^{\mathrm{D}} &= \frac{1}{n}\sum_i p_{ij}^{\mathrm{D}}, \qquad\qquad\qquad\qquad\quad \pi_j^{\mathrm{S}} = \frac{1}{K}\sum_k p_{jk}^{\mathrm{S}},
\end{aligned}$$

where ψ is the digamma function, i.e. the derivative of the log-Gamma function. Recalling that, for a variable $U \sim \Gamma(\alpha_1, \alpha_2)$, $\mathbb{E}[U] = \alpha_1/\alpha_2$ and $\mathbb{E}[\log U] = \psi(\alpha_1) - \log\alpha_2$, the update rule for the Gamma prior parameters on U_{ik} corresponds to averaging the moments and log-moments of the variational distribution on U_{ik} over i (similarly for V_{jk} over j). Regarding the Bernoulli prior parameters π_j^{D}, the update rule is also an average of the corresponding variational parameter over i (similarly for π_j^{S} over k).

A.3 Data generation

We set the hyper-parameters $(\alpha_{k,1}, \alpha_{k,2})_k$ and $(\beta_{k,1}, \beta_{k,2})_k$ of the Gamma prior distributions on U_{ik} and V_{jk} to generate structure in the data, i.e. groups of individuals and groups of variables.

Generation of U. In practice, individuals $i = 1, \ldots, n$ are partitioned into N balanced groups, denoted by $\mathcal{U}_1, \ldots, \mathcal{U}_N$. To do so, we generate a matrix $\mathbf{U}$ with blocks on the diagonal. Each block, denoted by $\mathcal{B}_{\mathbf{U},g}$ contains n/N rows and K/N columns. Each entry U_{ik} in each block $\mathcal{B}_{\mathbf{U},g}$ $(g = 1, \ldots, N)$ is drawn from a Gamma distribution $\Gamma(\alpha + \varepsilon_\alpha, 1)$ with a shape parameter depending on $\alpha > 0$ and an additive term $\varepsilon_\alpha > 0$. All entries U_{ik} that are not in the diagonal blocks of $\mathbf{U}$ are drawn from a Gamma distribution $\Gamma(\alpha, 1)$. Hence, each groups of individuals $\mathcal{U}_g$ corresponds to a block $\mathcal{B}_{\mathbf{U},g}$. Thus, this generation pattern requires that $K > N$. In addition, the term $\varepsilon_\alpha > 0$ quantifies how much the groups of individuals are distinct. In practice, we fix $\alpha = 4$, we use $\varepsilon_\alpha = 4$ or 8 (for low or high separation respectively) and $N = 2$ or 3 groups of individuals.

Generation of V. The question of simulating data based on a sparse representation $\mathbf{V}$ of the variables in our context of matrix factorization is not straightforward. Indeed, if we impose that some variables j do not contribute to any component k, i.e. that V_{jk} is null for any k, then $\sum_k U_{ik} V_{jk}$ is always null for $i = 1, \ldots, n$. Thus, the recorded data entry X_{ij} will be deterministic and null for any observation i (i.e. the j^{th} column in $\mathbf{X}$ will be null). There is no interest to generate full columns of null values in the matrix $\mathbf{X}$, since it is unnecessary to use a statistical analysis to determine that a column of zeros will not be informative. This question is not an issue about the formulation of the model, but rather concerns the generation of non informative columns in $\mathbf{X}$ that

will correspond to null rows in the matrix $\mathbf{V}$.

To overcome this issue, we use the following generative process. The variables $j = 1, \ldots, p$ are first partitioned into two groups $\mathcal{V}_0$ and $\mathcal{V}_\varnothing$ of respective sizes p_0 and $p - p_0$ (with $p_0 \leq p$). The p_0 variables in $\mathcal{V}_0$ will represent the pertinent variables for the lower dimensional representation, whereas variables in $\mathcal{V}_\varnothing$ will be considered irrelevant or noise. The matrix $\mathbf{V}$ will be a concatenation of two matrices $\mathbf{V}^0$ and $\mathbf{V}^\varnothing$:

$$\mathbf{V}_{p\times K} = \begin{pmatrix} \mathbf{V}^0 \\ \mathbf{V}^\varnothing \end{pmatrix}$$

All V_{jk} in $\mathbf{V}^\varnothing$ are drawn from a Gamma distribution $\Gamma(0.7, 1)$, so that $\mathbb{E}[V_{jk}]$ will be small but non null to avoid null columns in $\mathbf{X}$. The ratio p_0/p sets the expected degree of sparsity in the model. In practice, we set $p_0/p = 1, 0.6$ or 0.3 corresponding to different proportions of noisy genes ($0, 40$ or 70% of noisy genes).

To simulate dependency between recorded variables, we generate groups of variables in the set $\mathcal{V}_0$ of pertinent variables. We use a similar strategy as the one used to simulate $\mathbf{U}$. $\mathcal{V}_0$ is partitioned into P balanced groups, denoted by $\mathcal{V}_1, \ldots, \mathcal{V}_P$. We generate the corresponding matrix $\mathbf{V}^0$ with blocks on the diagonal. Each block, denoted by $\mathcal{B}_{\mathbf{V},g}$ contains p_0/P rows and K/P columns: Each entry V_{jk} in each block $\mathcal{B}_{\mathbf{V},g}$ $(g = 1, \ldots, P)$ is drawn from a Gamma distribution $\Gamma(\beta + \varepsilon_\beta, 1)$ with a shape parameter depending on $\beta > 0$ and an additive term $\varepsilon_\beta > 0$. All entries V_{jk} that are not in the blocks on diagonal are drawn from a Gamma distribution $\Gamma(\beta, 1)$. Hence, each groups of individuals $\mathcal{V}_g$ corresponds to a block $\mathcal{B}_{\mathbf{V},g}$. Again, this generation pattern requires that $K > N$. In addition, the term $\varepsilon_\beta > 0$ quantifies how much the groups of genes are distinct. In practice, we fix $\beta = 4$, we use $\varepsilon_\beta = 4$ or 8 (for high or low dependency respectively) and $P = 2$ groups of variables.

Generation of X. The data are simulated according to their conditional Poisson distribution in the model i.e. $\mathscr{P}(\sum_k u_{ik}\, v_{jk})$. In practice, we want to consider zero-inflation in the model, thus we consider the Dirac-Poisson mixture and simulate X_{ij} according to the following conditional distribution:

$$X_{ij} \,|\, (U_{ik}, V_{jk})_k, D_{ij} \sim (1 - D_{ij}) \times \delta_0 + D_{ij} \times \mathscr{P}(\textstyle\sum_k U_{ik}\, V_{jk})\,,$$

where the dropout indicator D_{ij} is drawn from a Bernoulli distribution $\mathcal{B}(\pi_j^{\mathrm{D}})$, the proportion of dropout events is set by the probability π_j^{D}. To generate data without dropout events, we just have to set $D_{ij} = 1$ for any couple (i, j), i.e. $\pi_j^{\mathrm{D}} = 1$ for any j.

In practice, we fix $K = 50$, $n = 100$ and $p = 1000$ to simulate our data. We generate different level of zero-inflation: $\pi_j^{\mathrm{D}} = 1$ for any j, corresponding to "no zero-inflation"; $\pi_j^{\mathrm{D}} \in [0.4; 0.6]$ corresponding to what we call "low zero-inflation"; and $\pi_j^{\mathrm{D}} \in [0.2; 0.4]$ corresponding to what we call "high zero-inflation" in the data.

A.4 Additional results

A.4.1 Computation time

Figure A.1 shows average computation time for the different methods (pCMF, Poisson-NMF, SPCA, ZIFA) for a single run on a single-core standard CPU with frequency between 2 and 2.5 GHz. All methods, including ours, have different levels of multi-threading and can benefit from multi-core CPU computations. We restrained to a single CPU core for each method run, because we were simultaneously running a huge number of simulations on a CPU cluster.

Our method shows comparable computation time as state-of-the-art approaches as Poisson-NMF (from the `NMF` R-package) or ZIFA (from the `ZIFA` Python-package). The sparse PCA (from the `PMA` R-package) is the gold standard regarding running time thanks to the efficiency of the PCA algorithm based on the Singular Value Decomposition (SVD) algorithm. However, we recall that (sparse) PCA shows poor results regarding clustering and data visualization.

Eventually, we mention that our method is available in an R-package, however our algorithms are implemented in interfaced C++ for computational efficiency.

A.4.2 Standard GaP versus our ZI sparse GaP factor model

Figure A.2 illustrates the interest of our zero-inflated sparse Gamma-Poisson factor model compared to the standard Gamma-Poisson factor model, especially in presence of dropout events and noisy genes. Our method pCMF based on our ZI sparse GaP factor model performs as well as the pCMF based on the standard GaP factor model when there is no dropout events in the data, independently from the proportion of noisy genes. In addition, when the level of zero-inflation is higher, we can see that the ZI-specific model outperforms the standard ones, highlighting the interest of our approach.

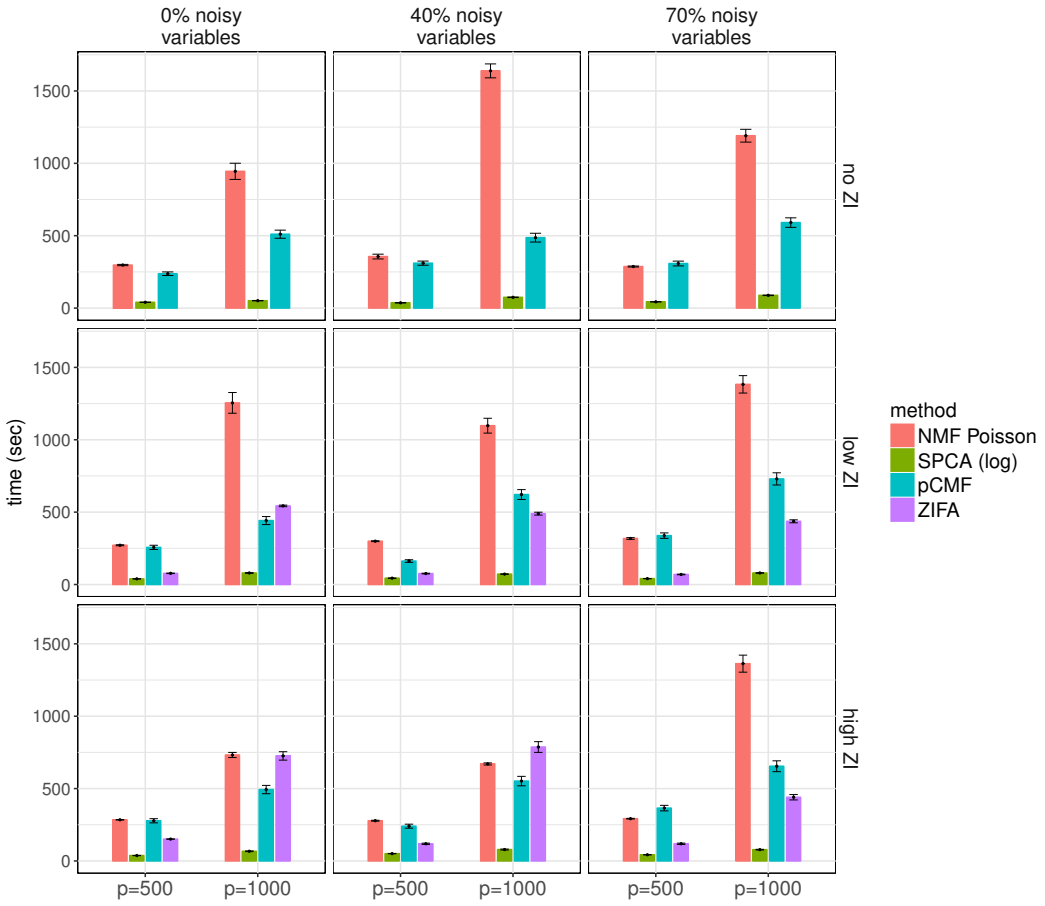

Figure A.1: Computation time on a single CPU core for the different approaches, depending on the number of variables p, for different levels of zero-inflation and different proportion of noisy variables in the data. The number of components is set to $K = 10$. Data are generated with $n = 100$ and 2 groups of individuals. Average values and deviation are estimated across 100 repetitions.

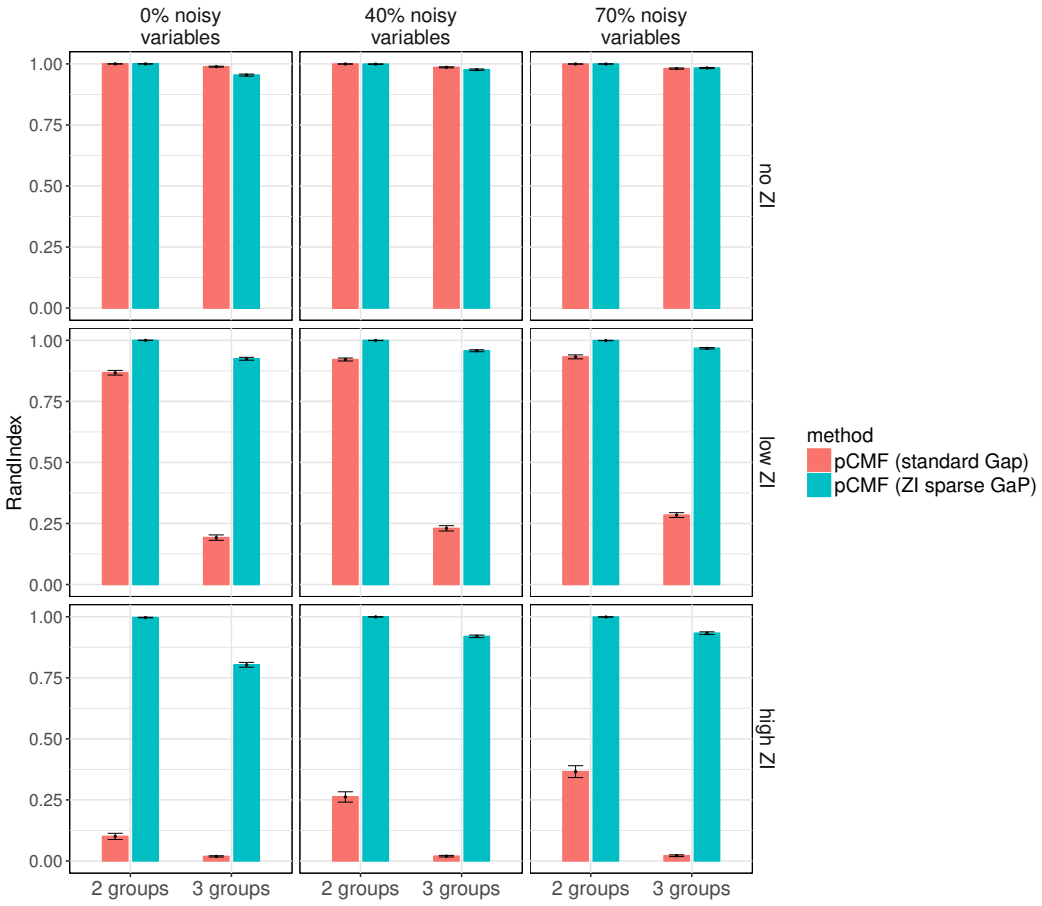

Figure A.2: Adjusted Rand Index comparing clusters found by a κ-means algorithm (applied to $\widehat{\mathbf{U}}$ with $\kappa = 2$) and the original groups of individuals, depending on the number of individual groups in the data, for different levels of zero-inflation and different proportion of noisy variables in the data. The number of components is set to $K = 10$. Data are generated with $n = 100$, $p = 1000$. Average values and deviation are estimated across 100 repetitions.