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 [1,2,3,*], L. Modolo [1,4,5], J. E. Mold [5], S. Lambert-Lacroix [6] and F. Picard [1]

March 12, 2019

[1]Univ Lyon, Université Lyon 1, CNRS, LBBE UMR 5558, Villeurbanne, France
[2]Univ Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK UMR 5224, Grenoble, France,
[3]Université de Montpellier, CNRS, IMAG UMR 5149, Montpellier, France,
[4]Univ Lyon, ENS Lyon, Université Lyon 1, CNRS, LBMC UMR 5239, Lyon, France,
[5]Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden,
[6]Univ Grenoble Alpes, CNRS, TIMC-IMAG UMR 5525, Grenoble, France.
*Corresponding author: ghislain.durif@umontpellier.fr

Abstract

Motivation: The development of high throughput single-cell sequencing technologies now allows the investigation of the population diversity of cellular transcriptomes. The expression dynamics (gene-to-gene variability) can be quantified more accurately, thanks to the measurement of lowly-expressed genes. In addition, the cell-to-cell variability is high, with a low proportion of cells expressing the same genes at the same time/level. Those emerging patterns appear to be very challenging from the statistical point of view, especially to represent a summarized view of single-cell expression data. PCA is a most powerful tool for high dimensional data representation, by searching for latent directions catching the most variability in the data. Unfortunately, classical PCA is based on Euclidean distance and projections that poorly work in presence of over-dispersed count data with dropout events like single-cell expression data.

Results: 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. It is able to jointly build a low dimensional representation of cells and genes. We show how this probabilistic framework induces a geometry that is suitable for single-cell data visualization, and produces a compression of the data that is very powerful for clustering purposes. Our method is competed against other standard representation methods like t-SNE, and we illustrate its performance for the representation of single-cell expression (scRNA-seq) data.

Availability: Our work is implemented in the pCMF R-package[1].

[1] https://github.com/gdurif/pCMF

1 Introduction

The combination of massive parallel sequencing with high-throughput cell biology technologies has given rise to the field of single-cell genomics, which refers 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 heterogeneities at the cellular level (Deng *et al.*, 2014; Saliba *et al.*, 2014). 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 intestine cells (Grün *et al.*, 2015). The construction of a comprehensive atlas of human cell types is now within our reach (Wagner *et al.*, 2016). The characterization of heterogeneities in single-cell expression data thus requires an appropriate statistical model, as 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; Riggs and Lalonde, 2017, Chap. 6). 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).

Principal component analysis (PCA) is one of the most widely used dimension reduction technique, as it allows the quantification and visualization of variability in massive datasets. It consists in approximating the observation matrix $\mathbf{X}_{[n\times m]}$ (n cells, m genes), by a factorized matrix of reduced rank, denoted $\mathbf{U}\mathbf{V}^T$ where $\mathbf{U}_{[n\times K]}$ and $\mathbf{V}_{[m\times K]}$ represent the latent structure in the observation and variable spaces respectively. This projection onto a lower-dimensional space (of dimension K) allows one to catch gene co-expression patterns and clusters of individuals. PCA can be viewed either geometrically or through the light of a statistical model (Landgraf and Lee, 2015). Standard PCA is based on the ℓ_2 distance as a metric and is implicitly based on a Gaussian distribution (Eckart and Young, 1936). Model-based PCA offers the unique advantage to be adapted to the data distribution. It consists in specifying the distribution of the data $\mathbf{X}_{[n\times m]}$ through a statistical model, and to factorize $\mathbb{E}(\mathbf{X})$ instead of $\mathbf{X}$. In this context the ℓ_2 metric is replaced by the Bregman divergence which is adapted to maximum likelihood inference (Collins *et al.*, 2001). A probabilistic version of the Gaussian PCA was proposed by Pierson and Yau (2015) in the context of single cell data analysis, with the modeling of zero inflation (the ZIFA method). ScRNA-seq data may be better analyzed by methods dedicated to count data such as the Non-negative Matrix Factorization (NMF) introduced in a Poisson-based framework by Lee and Seung (1999) or the Gamma-Poisson factor model (Cemgil, 2009; Févotte and Cemgil, 2009; Landgraf and Lee, 2015). None of the currently available dimension reduction

methods fully model single-cell expression data, characterized by over-dispersed zero inflated counts (Kharchenko *et al.*, 2014; 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. In particular, gene expression can be normalized but does not require to be transformed (log, Anscombe) in our framework. 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 (Simchowitz, 2013), and we introduce sparsity in the model thanks to a spike-and-slab approach (Malsiner-Walli and Wagner, 2011) that is based on a two component sparsity-inducing prior on loadings (Titsias and Lázaro-Gredilla, 2011). We propose a heuristic to initialize the sparsity layer based on the variance of the recorded variables, acting as an integrated gene filtering step, which is an important issue in scRNA-seq data analysis (Soneson and Robinson, 2018). The model is inferred using a variational EM algorithm that scales favorably to data dimension compared with Markov Chain Monte-Carlo (MCMC) methods (Hoffman *et al.*, 2013; 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, following a strategy that is standard in the Generalized Linear Models framework. Moreover, we show that our criterion corresponds to the percentage of explained variance in the PCA case, which makes it suitable to compare geometric and probabilistic methods.

We show the performance of pCMF on simulated and experimental datasets, in terms of visualization and quality of fit. Moreover, we show the benefits of using pCMF as a preliminary dimension reduction step before clustering or before the popular t-SNE approach (van der Maaten and Hinton, 2008; Amir *et al.*, 2013). Experimental published data are used to show the capacity of pCMF to provide a better representation of the heterogeneities within scRNA-Seq datasets, which appears to be extremely helpful to characterize cell types. Finally, our approach also provides a lower space representation for genes (and not only for cells), contrary to t-SNE. pCMF is implemented in the form of a R package available at `https://github.com/gdurif/pCMF`.

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

The Poisson factor model. Our data consist of a matrix of counts (potentially normalized but not transformed), denoted by $\mathbf{X} \in \mathbb{N}^{n \times m}$, that we want to decompose onto a subspace of dimension K (being fixed). In a first step we suppose that the data follow a multivariate Poisson distribution of intensity $\mathbf{\Lambda}$. Following the standard Poisson Non-Negative Matrix Factorization (Poisson NMF, Lee and Seung, 1999), we

approximate this intensity such that

$$\mathbf{X} \sim \mathscr{P}\big(\mathbf{\Lambda}\big), \quad \mathbf{\Lambda} \simeq \mathbf{U}\mathbf{V}^T, \tag{1}$$

with factor $\mathbf{U} \in \mathbb{R}^{+,n\times K}$ the coordinates of the n observations (cells) in the subspace of dimension K, and loadings $\mathbf{V} \in \mathbb{R}^{+,m\times K}$ the contributions of the m variables (genes).

Modeling over-dispersion. We account for over-dispersion by using the the Negative-Binomial distribution (Anders and Huber, 2010), through a hierarchical Gamma-Poisson representation (GaP) Cemgil (2009). In our factor model $\mathbf{U}$ and $\mathbf{V}$ are modeled as independent random latent variables with Gamma distributions such that

$$\begin{aligned} U_{ik} &\sim \Gamma(\alpha_{k,1}, \alpha_{k,2}) \text{ for any } (i,k) \in [1:n] \times [1:K]\,, \\ V_{jk} &\sim \Gamma(\beta_{k,1}, \beta_{k,2}) \text{ for any } (j,k) \in [1:m] \times [1:K]\,. \end{aligned} \tag{2}$$

In practice, some third-party latent variables are introduced for the derivation of our inference algorithm (Cemgil, 2009; Zhou *et al.*, 2012). We consider latent variables $\mathbf{Z} = [Z_{ijk}] \in \mathbb{R}^{n\times m\times K}$, defined such that $X_{ij} = \sum_k Z_{ijk}$. These new indicator variables quantify 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.

Dropout modeling with a zero-inflated (ZI) model. To model zero-inflation, i.e. random null observations called dropout events, we introduce a dropout indicator variable $D_{ij} \in \{0, 1\}$ for $i = 1, \ldots, n$ and $j = 1, \ldots, p$ (c.f. Simchowitz, 2013). In this context, each $D_{ij} = 0$ if gene j has been subject to a dropout event in cell i, with $D_{ij} \sim \mathcal{B}(\pi_j^D)$. We consider gene-specific dropout rates, π_j^D, following recommendations of the literature (Pierson and Yau, 2015). Thus, to include zero-inflation in the probabilistic factor model, we consider that

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

where δ_0 is the Dirac mass at 0, i.e. $\delta_0(X_{ij}) = 1$ if $X_{ij} = 0$ and 0 otherwise. The dropout indicators D_{ij} are assumed to be independent from the factors U_{ik} and V_{jk}. Then, by integrating D_{ij} out, the probability of observing a zero in the data illustrates the two potential sources of zeros and becomes

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

Thus, inference will be based on the factors U_{ik} and V_{jk} and on probabilities π_j^D.

Probabilistic variable selection. Finally we suppose that our model is parsimonious. We consider that among all recorded variables, only a proportion carries the signal and the others are 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 (Engelhardt and Adams, 2014). The model is then enriched by the introduction of a new indicator variable $S_{jk} \sim \mathcal{B}(\pi_j^S)$, that equals 1 if gene j contributes to 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 (Mitchell and 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). The contribution of gene j to the component k is accounted for in the conditional Poisson distribution of X_{ij}, with

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

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

Underlying geometry. Knowing $\mathbf{U}$ and $\mathbf{V}$, to quantify the approximation of matrix $\mathbf{X}$ by $\mathbf{U}\mathbf{V}^T$, we consider the Bregman divergence, that can be viewed as a generalization of the Euclidean metric to the exponential family (see Collins *et al.*, 2001; Banerjee *et al.*, 2005; Chen *et al.*, 2008). In the Poisson model, the Bregman divergence between $\mathbf{X}$ and $\mathbf{U}\mathbf{V}^T$ is defined as (Févotte and Cemgil, 2009):

$$D(\mathbf{X}\,|\,\mathbf{U}\mathbf{V}^T) = \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}\,\log\left(\frac{x_{ij}}{\sum_k U_{ik}V_{jk}}\right) - x_{ij} + \sum_k U_{ik}V_{jk}.$$

Hence the geometry is induced by an appropriate probabilistic model dedicated to count data. Potential identifiability issues are addressed in Supp. Mat. (Section S.2).

In the following, we will refer to pCMF for the method implementing the model with dropout but the without sparsity layer, and to sparse pCMF (or spCMF) for the model with dropout and sparsity layers.

2.1 Quality of the reconstruction.

The Bregman divergence between the data matrix $\mathbf{X}$ and the reconstructed matrix $\widehat{\mathbf{U}}\widehat{\mathbf{V}}^T$ in our GaP factor model is related to the deviance of the Poisson model defined such as (Landgraf and Lee, 2015)

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

where $\log p(\mathbf{X} \mid \mathbf{\Lambda})$ is the Poisson log-likelihood based on the matrix notation (1). We have $\text{Dev}(\mathbf{X}, \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T) \propto D(\mathbf{X} \mid \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T)$, thus the deviance can be used 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. We introduce a criterion that we call percentage of explained deviance that is a generalization of the percentage of explained variance to our GaP factor model. However, since our models are not nested for increasing K, the definition of this criterion appears non trivial. To assess the quality of our model, we propose to define the percentage of explained deviance as:

$$\%\text{dev} = \frac{\log p(\mathbf{X} \mid \mathbf{\Lambda} = \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T) - \log p(\mathbf{X} \mid \mathbf{\Lambda} = \mathbf{1}_n\bar{\mathbf{X}})}{\log p(\mathbf{X} \mid \mathbf{\Lambda} = \mathbf{X}) - \log p(\mathbf{X} \mid \mathbf{\Lambda} = \mathbf{1}_n\bar{\mathbf{X}})} \tag{3}$$

where $\widehat{\mathbf{U}}\widehat{\mathbf{V}}^T$ is the predicted reconstructed matrix in our model, $\mathbf{1}_n$ is a column vector filled with 1 and $\bar{\mathbf{X}}$ is a row vector of size m storing the column-wise average of $\mathbf{X}$. We use two baselines: (i) the log-likelihood of the saturated model, i.e. $\log p(\mathbf{X} \mid \mathbf{\Lambda} = \mathbf{X})$ (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. $\log p(\mathbf{X} \mid \mathbf{\Lambda} = \mathbf{1}_n\bar{\mathbf{X}})$, which is the most simple model that we could use. This formulation ensures that the ratio %dev lies in $[0; 1]$. An interesting point is that if we assume a Gaussian distribution on the data, the percentage of explained deviance is exactly the percentage of explained variance from PCA (c.f. Section S.1), which makes our criterion suitable for to compare different factor models.

2.2 Choosing the dimension of the latent space

As noticed in the previous section, the GaP factor model with an increasing number K of factors are not nested (the model associated to the NMF presents the same properties). Consequently, testing different values of K requires to fit different models (contrary to PCA for instance). We choose the number of factors by fitting a model with a large K and verifying how the matrix $\widehat{\mathbf{U}}_{1:k}(\widehat{\mathbf{V}}_{1:k})^T$ reconstructs $\mathbf{X}$ depending on $k = 1, \ldots, K$ with a rule of thumb based on the "elbow" shape of the fitting criterion . This approach is for instance widely used in PCA by checking the proportion of variability explained by each components, see Friguet (2010, p.96) for a review of the different criteria to choose K in this context. Here we use the deviance, or equivalently the Bregman divergence $k \mapsto D\big(\mathbf{X} \mid \widehat{\mathbf{U}}_{1:k}(\widehat{\mathbf{V}}_{1:k})^T\big)$ to find the latent dimension from where adding new factors does not improve $D\big(\mathbf{X} \mid \widehat{\mathbf{U}}_{1:k}(\widehat{\mathbf{V}}_{1:k})^T\big)$. This determination is however not always unambiguous and may sometimes lead to some over-fitting, i.e. when considering too much factors. In addition, when focusing on data visualization, we generally set $K = 2$.

3 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 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). In particular, we extend the version of the variational EM algorithm (Beal and Ghahramani, 2003) proposed by Dikmen and Févotte (2012) in the context of the standard Gamma-Poisson factor model to our sparse and zero-inflated GaP model. Figure S.1 in Supp. Mat. gives an overview of the variational framework.

3.1 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:

$$J(q) = \mathbb{E}_q[\log p(\mathbf{X}, \mathbf{Z}, \mathbf{U}, \mathbf{V}', \mathbf{S}, \mathbf{D})] - \mathbb{E}_q[\log q(\mathbf{Z}, \mathbf{U}, \mathbf{V}', \mathbf{S}, \mathbf{D})]\,, \tag{4}$$

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:

$$\begin{aligned} q(\mathbf{Z}, \mathbf{U}, \mathbf{V}', \mathbf{S}, \mathbf{D}) &= \prod_{i=1}^{n}\prod_{j=1}^{m} q\big((Z_{ijk})_k \,|\, (r_{ijk})_k\big) \\ &\times \prod_{i=1}^{n}\prod_{k=1}^{K} q(U_{ik} \,|\, \mathbf{a}_{ik}) \times \prod_{j=1}^{m}\prod_{k=1}^{K} q(V'_{jk} \,|\, \mathbf{b}_{jk}) \\ &\times \prod_{j=1}^{m}\prod_{k=1}^{K} q(S_{jk} \,|\, p^S_{jk}) \times \prod_{i=1}^{n}\prod_{j=1}^{m} q(D_{ij} \,|\, p^D_{ij}) \end{aligned} \tag{5}$$

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

3.2 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:

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

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}).

3.3 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, e.g. $p(U_{ik}\,|\,—)$, $p(V_{jk}\,|\,—)$ and $p\big((Z_{ijk})_k\,|\,—\big)$. The partial formulation are therefore respectively:

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

Similar formulations can be derived regarding parameters p^D_{ij} and p^S_{jk}. Therefore, the ELBO is explicit regarding each variational parameter and the gradient of the ELBO $J(q)$ depending on the variational parameters $\mathbf{a}_{ik}$, $\mathbf{b}_{jk}$, r_{ijk}, p^D_{ij} and p^S_{jk} respectively can be derived to find the coordinate of the stationary point (corresponding to a local optimum). In our factor model all full conditionals are tractable (c.f. Section S.4.1 in Supp. Mat.). In practice, thanks to the formulation in the exponential family, the optimum value for each variational parameter corresponds to the expectation regarding q of the corresponding parameter of the full conditional distribution (see Hoffman

et al., 2013). Thus the coordinates of the ELBO's gradient optimal point are explicit. We mentioned that distributions with a mass at 0 (zero-inflated Poisson or sparse Gamma) lie in the exponential family (Eggers, 2015) and the general formulation from Hoffman *et al.* (2013) remains valid. Detailed formulations of update rules regarding all variational parameters are given in Supp. Mat. (Section S.4.2).

3.4 Variational EM algorithm

We use the variational-EM algorithm (Beal and Ghahramani, 2003) to jointly approximate the posterior distributions and to estimate the hyper-parameters $\mathbf{\Omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\pi}^S, \boldsymbol{\pi}^D)$. 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 $\mathbb{E}[p(\mathbf{X}, \mathbf{U}, \mathbf{V}', \mathbf{S}, \mathbf{D}\,;\,\mathbf{\Omega})|\mathbf{X}]$ is approximated by $\mathbb{E}_q[p(\mathbf{X}, \mathbf{U}, \mathbf{V}', \mathbf{S}, \mathbf{D}\,;\,\mathbf{\Omega})]$. Then the variational M-step consists in maximizing $\mathbb{E}_q[p(\mathbf{X}, \mathbf{U}, \mathbf{V}', \mathbf{S}, \mathbf{D}\,;\,\mathbf{\Omega})]$ w.r.t. the hyper-parameters $\mathbf{\Omega}$. 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. Detailed formulations of update rules regarding all prior hyper-parameters are given in Supp. Mat. (Section S.4.3).

3.5 Initialization of the algorithm

To initialize variational and hyper-parameters of the model, we sample $\mathbf{U}$ and $\mathbf{V}$ from Gamma distributions such that $X_{ij} \simeq \sum_k U_{ik} V_{jk}$ on average. The Gamma variational and hyper parameters are initialized from these values following the update rules detailed in Supp. Mat. Section S.4.2. Dropout probabilities p_{ij}^{D} and π_j^{D} are initialized by $1/n \sum_i \mathbf{1}_{\{X_{ij}>0\}}$, i.e. the proportion of non-zero in the data for the corresponding gene. To initialize the sparsity probabilities p_{jk}^{S} and π_j^{S}, we use a heuristic based on the variance of the corresponding gene j. Assuming that genes with low variability will have less impact on the structure embedded in the data, we propose a starting value such that

$$P_j^{(0)} = 1 - \exp\left(-\widehat{s}_j / \widehat{m}_j\right) , \tag{6}$$

where $\widehat{m}_j$ is the mean of the non-null observations for gene j and $\widehat{s}_j$ its standard deviation (including null values). The scaling is better when removing the null values to compute the mean. This quantity adapts to the empirical variance of the observations, and will be close to 0 for genes with low variance, and close to 1 for genes with high variability.

4 Empirical study of pCMF

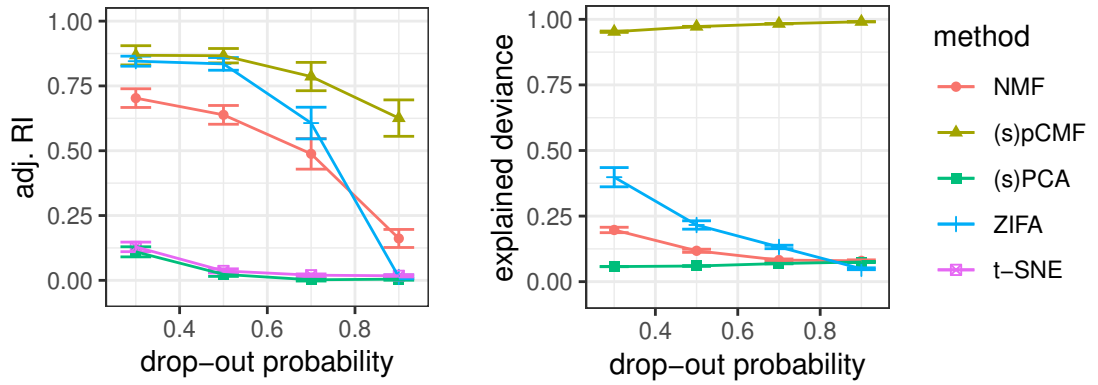
All codes are available on a public repository for reproducibility[2]. We compare our method with standard approaches for unsupervised dimension reduction: the Poisson-NMF (Lee and Seung, 1999), applied to raw counts (model-based matrix factorization approach based on the Poisson distribution); the PCA (Pearson, 1901) and the sparse PCA (Witten *et al.*, 2009), based on an ℓ_1 penalty in the optimization problem defining the PCA to induce sparsity in the loadings $\mathbf{V}$, both applied to log counts. We use sparse methods (sparse PCA, sparse pCMF) with a re-estimation step on the selected genes. We will refer to them as (s)PCA and (s)pCMF in the results respectively, to differentiate them from sparse PCA and sparse pCMF (without re-estimation), PCA and pCMF (without the sparsity layer). In addition, we use the Zero-Inflated Factor Analysis (ZIFA) by Pierson and 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. We also compare our method with t-SNE that is commonly used for data visualization (van der Maaten and Hinton, 2008). It requires to choose a "perplexity" hyper-parameter that cannot be automatically calibrated, thus being less appropriate for a quantitative analysis. In the following, we always choose the perplexity values that gives the better clustering results.

4.1 Simulated data analysis

To generate synthetic data we follow the hierarchical Gamma-Poisson framework as adopted by others (Zappia *et al.*, 2017). Details are provided in the Supp. Mat. (Section S.5). We generate synthetic multivariate over-dispersed counts, with $n = 100$ individuals and $m = 800$ recorded variables. We artificially create clusters of individuals (with different level of dispersion) and groups of dependent variables. We 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 m 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, the range of our simulation parameters being comparable to other published simulated data (Zappia *et al.*, 2017).

We train the different methods with $K = 2$ to visualize the reconstructed matrices $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ (c.f. Section 2). To assess the ability of each method to retrieve the structure of cells or genes, we run a κ-mean clustering algorithm on $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$ respectively (with $\kappa = 2$) and we measure the adjusted Rand Index (Rand, 1971) quantifying

[2] https://github.com/gdurif/pCMF_experiments



(a) (b)

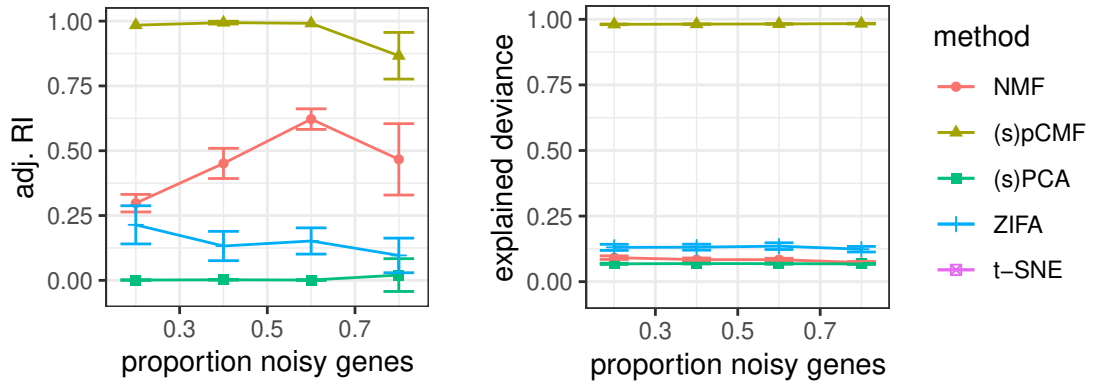
Figure 1: Adjusted Rand Index (1a) for the clustering on $\widehat{\mathbf{U}}$ versus the true groups of cells; and explained deviance (1b) depending on the probability used to generate dropout events. Average values and deviation are estimated across 50 repetitions.

the accordance between the predicted clusters and original groups of cells or genes. Regarding our approach pCMF, we use $\log \widehat{\mathbf{U}}$ and $\log \widehat{\mathbf{V}}$ for data visualization and clustering because the log is the canonical link function for Gamma models. In addition, we also compute the percentage of explained deviance associated to the model to assess the quality of the reconstruction. Regarding the PCA (sparse or not) and ZIFA, we use the standard explained variance criterion (c.f. Section 2.1).

4.1.1 Clustering in the observation space

Effect of zero-inflation and cell representation. We first assess the robustness of the different methods to the level of dropout or zero-inflation (ZI) in the data. We generate data with 3 groups of observations (c.f. Supp. Mat. Section S.5) with a wide range of dropout probabilities. Figure 1a shows that (s)pCMF adapts to the level of dropout in the data and recovers the original clusters (high adjusted Rand Index) even with dropouts. Despite comparable performance with low dropout, Poisson-NMF and ZIFA are very sensitive to the addition of zeros. In addition, methods based on transformed counts like (s)PCA and t-SNE perform poorly, as they do not account for the specificity of the data (discrete, over-dispersed, (O'Hara and Kotze, 2010)).

Effect of noisy genes and gene representation. To quantify the impact of noisy genes on the retrieval of the clusters, we consider data generated with different proportions of noisy genes (genes that do not induce any structure in the data). We generate data with three groups of genes: two groups inducing some latent structure and one group of noisy genes (c.f. Supp. Mat. Section S.5). Figure 2 shows that (s)pCMF



(a) (b)

Figure 2: Adjusted Rand Index (2a) for the clustering on $\widehat{\mathbf{V}}$ versus the true groups of genes; and explained deviance (2b) depending on the proportion of noisy genes. Average values and deviation are estimated across 50 repetitions.

identifies correctly the clusters of genes, including the set of noisy genes, contrary to other approaches. This point shows that our approach correctly identifies the genes that support the lower-dimensional representation.

In addition to the clustering results, we compared the selection accuracy of the only two methods (sPCA, spCMF) that perform variable selection (Supp. Mat. Figure S.2). A selected gene is a gene that contributes to any latent dimension (any $V_{jk} \neq 0$). Sparse pCMF performs better than sparse PCA for various latent dimension K even for high levels of noisy genes. Sparse PCA shows better selection accuracy when the proportion of noisy genes is low. This point suggests that sparse pCMF would be less sensitive to gene pre-filtering when analysing scRNA-seq data, which corresponds to a removal of noisy genes and is generally crucial (Soneson and Robinson, 2018).

Details about data generation and additional data configuration regarding Figures 1 and 2 can be found in Supp. Mat. (Section S.7, especially Figures S.3 and S.4).

4.1.2 Data visualization

Data visualization is central in single-cell transcriptomics for the representation of high dimensional data in a lower dimensional space, in order to identify groups of cells, or to illustrate the cells diversity (e.g. Llorens-Bobadilla *et al.*, 2015; Segerstolpe *et al.*, 2016). In the matrix factorization framework, we represent observation (cell) coordinates and variable (gene) contributions: resp. $(\widehat{u}_{i1}, \widehat{u}_{i2})_{i=1,\dots,n}$ and $(\widehat{v}_{i1}, \widehat{v}_{i2})_{i=1,\dots,n}$ (or their log transform for pCMF) when the dimension is $K = 2$ (see Section 2).

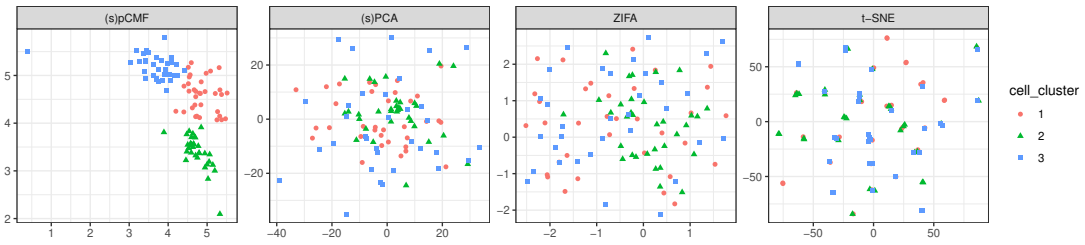


Figure 3: Representation of cells in a subspace of dimension $K = 2$. Here we have 60% of noisy variables, and dropout probabilities around 0.9.

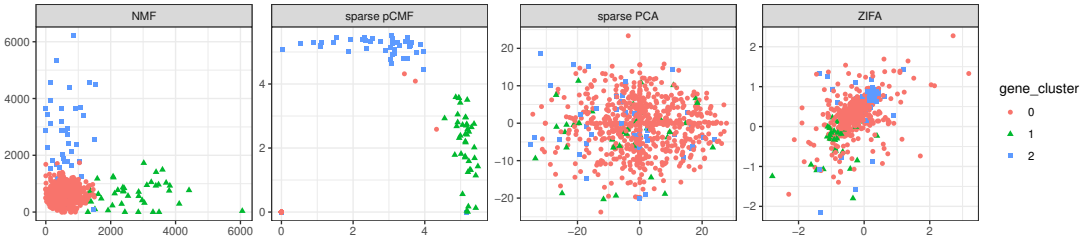


Figure 4: Representation of genes in a subspace of dimension $K = 2$ Here we have 80% of noisy variables, and dropout probabilities around 0.7. The label 0 corresponds to noisy genes.

We consider the same simulated data as previously ($n = 100$, $m = 800$, with three groups of cells, two groups of relevant genes and the set of noisy genes). Our visual results are consistent with the previous clustering results both regarding cell and gene visualization. In this challenging context (high zero-inflation and numerous noisy variables), by using pCMF, we are able to graphically identify the groups of individuals (cells) in the simulated zero-inflated count data (c.f. Figure 3). On the contrary, the 2-D visualization is not successful with PCA, ZIFA, Poisson-NMF and t-SNE, illustrating the interest of our data-specific approach. This point supports our claim that using data-specific model improves the quality of the reconstruction in the latent space.

In addition, linear projection methods (NMF, PCA, pCMF, ZIFA) can be used to visualize the contribution of genes to the principal axes (c.f. Figure 4). Thanks to sparsity constraints, the contribution of noisy genes are mostly set to 0 for sparse pCMF. Surprisingly, this selection is not efficient in the case of sparse PCA, indicating a lack of calibration of the sparsity constraint when data are counts. In comparison, Poisson NMF and ZIFA (not sparse) do not identity the cluster of noisy genes as clearly as sparse pCMF.

To quantify the model quality of the different methods on simulated data, we used the deviance associated to each method (c.f. Section 2.1). Figures 1b and 2b shows that the dimension reduction proposed by pCMF has excellent fit to the data regardless the level of dropout or the proportion of noisy genes, as compared with other methods.

Additional comparisons of computational time show that PCA is fastest method (but with low performance), whereas (sparse) pCMF is faster than ZIFA and sparse PCA, with increased performance (cf Supp. Mat. Section S.7).

4.2 Analysis of single-cell data

We now illustrate the performance of pCMF on different recent and large single-cell RNA-seq datasets that are publicly availbale: the Baron *et al.* (2016) dataset, the goldstandard and silverstandard datasets used in Freytag *et al.* (2018) (we used the silverstandard dataset 5 which was the largest). We also consider an older and smaller dataset from (Llorens-Bobadilla *et al.*, 2015) which is interesting because it discribes a continuum of activation in Neural Stem Cells (NSC). All datasets are available here[3] with the codes. More details about their origins are given in Supp. Mat. Section S.7.5. We consider large datasets with $\sim$ 1000 or $\sim$ 10000 cells to test the ability of our approach to face the expected increase of data volume in the next couple of years.

We present some quantitative results about clustering and data reconstruction (c.f. Table 1) and the corresponding qualitative results about cell visualization (c.f. Figures 5 and S.7 to S.9 in Supp. Mat.) and gene visualization (c.f. Figures S.10 to S.13 in Supp. Mat.). For each dataset (except the one from Llorens-Bobadilla *et al.*, 2015, where we used their pre-filtering), we use the same pipeline, we filter out genes expressed in less than 5% of the cells. In a second step, we remove genes for which the variance heuristic defined in Equation (6) is low. In practice we removed genes for which $P_j^{(0)} \leq 0.2$. Our idea was to remove uninformative genes, since pre-filtering is crucial (Soneson and Robinson, 2018) in such data, but also to reduce the number of genes to reduce the computation cost, in particular for ZIFA. Then we compare (s)pCMF, PCA, ZIFA and t-SNE. We discarded (s)PCA because the sparse PCA is computationally expensive (c.f. Supp. Mat. Section S.7.3) due to the required cross-validation.

A general empirical property is that clustering accuracy dicreases for all methods when the number of groups of cells increases. However, our approach (s)pCMF produces a better (or as good) view of the cells regarding clustering purpose in every examples, since the adjusted Rand Index is higher (c.f. Table 1). We observe the same trend regarding the quality of the reconstruction since the explained deviance is generally also higher for (s)pCMF. Data visualization is not always clear (c.f. Figures S.7 and S.8), especially when the number of groups is large as in Baron *et al.* (2016) or Freytag *et al.* (2018) silverstandard, however it is possible to clearly identify large clusters of cells in the (e.g. beta cells in Baron *et al.* (2016) or CD14+ Monocyte in Freytag *et al.* (2018) silverstandard) with our method (and some of the others). On the goldstandard dataset from Freytag *et al.* (2018), the difference regarding cells representation between the different approaches is more visual (c.f. Figure 5), where our approach is the

[3] https://github.com/gdurif/pCMF_experiments

only one that is able to distinctly identify the three populations of cells. On the Llorens-Bobadilla *et al.* (2015) dataset, our approach clearly highlights this continuum of activation presented in their paper, which can also be seen with ZIFA, but is not as much clear with PCA and t-SNE.

Regarding gene visualization (c.f. Figures S.10 to S.13 in Supp. Mat.), we compare the representation of sparse pCMF to PCA, ZIFA (and not t-SNE since it does not jointly learn $\mathbf{U}$ and $\mathbf{V}$). The interest of sparsity for gene selection is to highlight more precisely the genes that contribute to the latent representation. For each dataset, it is possible to detect which genes are important for each latent dimension: some are null on both (e.g. uninformative genes), some contribute to a single dimension, some contribute to both. This pattern is not as clear with methods that do not implement any sparsity layers.

These different points show the interest of our approach to analyze recent single-cell RNA-seq datasets, even large ones. Empirical properties studied on simulations are confirmed on experimental data: providing a dimension reduction method adapted to single cell data, where the sparsity constraints is powerful to represent complex single cell data. In addition, our heuristic to assume gene importance based on their variance appears to be efficient (i) to perform a rough pre-filtering to reduce the dimension and (ii) to discriminate between noisy genes and informative ones directly in the sparse pCMF algorithm. In addition, it appears that our method is fast compared to ZIFA for instance, since it takes less than two minutes on the different examples to run (s)pCMF (sparse pCMF + re-estimation), on a 16-core machine, whereas ZIFA can take between 5 and 25 minutes on the same architecture.

	nb cells	nb genes (before pre-filter.)	prop. 0	nb group		(s)pCMF	PCA	ZIFA	t-SNE
Baron *et al.* (2016)	1886	6080 (14878)	80.9%	13	adj. RI	**21.2**%	14.3%	15.4%	14.2%
					%dev	**73.2**%	41.6%	53.5%	/
Freytag *et al.* (2018) goldstandard	925	8580 (58302)	39.5%	3	adj. RI	**81.3**%	60.1%	56.8%	60.5%
					%dev	55.7%	**65.6**%	48.6%	/
Freytag *et al.* (2018) silverstandard 5	8352	4547 (33694)	86.3%	11	adj. RI	24.2%	16.2%	19.8%	**24.8**%
					%dev	**70.0**%	55.1%	/	/
Llorens-Bobadilla *et al.* (2015)	141	13826 (43309)	64.8%	6	adj. RI	**40.1**%	25.3%	38.3%	29.8%
					%dev	**64.4**%	34.8%	42.6%	/

Table 1: Performance of the different methods regarding quality of reconstruction (percentage of explained deviance) and clustering (adjusted Rand Index). Each scRNA-seq dataset is characterized by the number of cells, the number of genes used in the analysis (we specify the original number before the pre-filtering step) and by the number of original groups. The adjusted Rand Index compares clusters found by a κ-means algorithm (applied to $\widehat{\mathbf{U}}$ with $\kappa =$ nb group) and the original groups of cells.

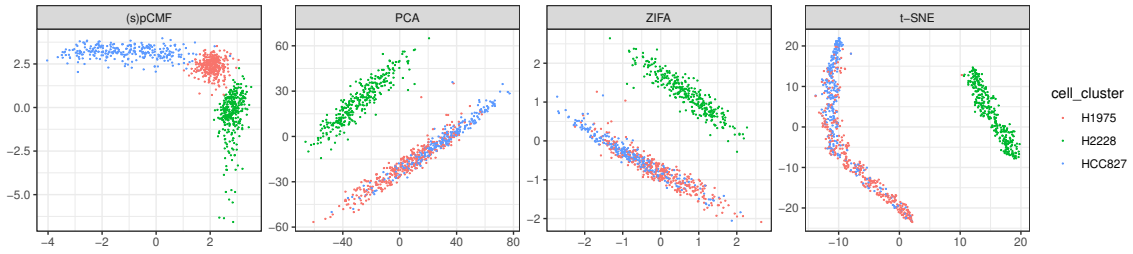


Figure 5: Analysis of the goldstandard scRNA-seq data from Freytag *et al.* (2018), 925 cells, 8580 genes. Visualization of the cells in a latent space of dimension 2.

5 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 analyze single-cell RNA-seq data. Matrix factorization allows to jointly construct a lower dimensional representation of cells and genes. Our probabilistic Count Matrix Factorization (pCMF) approach accounts for the specificity of these data, being zero-inflated and over-dispersed counts. In other words, 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 for characterizing 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 particular, pCMF data representation is less sensitive to the choice of the latent dimension K regarding clustering results, which supports the interest of our approach for data exploration. It appears (through the explained deviance criterion that we introduced) that the reconstruction learned by pCMF better represents the variability in the data (compared to PCA or ZIFA). In addition, pCMF can select genes that explain the latent structure in the data, thanks to a sparsity layer which does not require any parameter tuning.

An interesting direction to improve pCMF would be to integrate covariables or con-

founding factors in the Gamma-Poisson model, for instance to account for technical effect in the data or for data normalization. A similar framework based on zero-inflated Negative Binomial distribution was proposed by Risso *et al.* (2017), and it could be extended to our framework of matrix factorization

Funding

This work was supported by the french National Resarch Agency (ANR) as part of the "ABS4NGS" project [ANR-11-BINF-0001-06] and as part of the "MACARON" project [ANR-14-CE23-0003], and by the European Research Council as part of the ERC grant Solaris. It was performed using the computing facilities of the computing center LBBE/PRABI.

References

Amir, e.-A., Davis, K. L., Tadmor, M. D., Simonds, E. F., Levine, J. H., Bendall, S. C., Shenfeld, D. K., Krishnaswamy, S., Nolan, G. P., and 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.

Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, **11**(10), R106.

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

Baron, M., Veres, A., Wolock, S. L., Faust, A. L., Gaujoux, R., Vetere, A., Ryu, J. H., Wagner, B. K., Shen-Orr, S. S., Klein, A. M., Melton, D. A., and Yanai, I. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell systems*, **3**(4), 346–360.e4.

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

Buenrostro, J. D., Wu, B., Litzenburger, U. M., Ruff, D., Gonzales, M. L., Snyder, M. P., Chang, H. Y., and 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., and 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, H.-I. H., Jin, Y., Huang, Y., and Chen, Y. (2016). Detection of high variability in gene expression from single-cell RNA-seq profiling. *BMC Genomics*, **17**(Suppl 7).

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

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

Deng, Q., Ramsköld, D., Reinius, B., and Sandberg, R. (2014). Single-cell RNA-seq reveals dynamic, random monoallelic gene expression in mammalian cells. *Science*, **343**(6167), 193–196.

Dikmen, O. and 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. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**(3), 211–218.

Eggers, J. (2015). On Statistical Methods for Zero-Inflated Models. Technical Report U.U.D.M. Project Report 2015:9, Uppsala Universitet.

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

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

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, **97**(458), 611–631.

Freytag, S., Tian, L., Lönnstedt, I., Ng, M., and Bahlo, M. (2018). Comparison of clustering tools in R for medium-sized 10x Genomics single-cell RNA-sequencing data. *F1000Research*, **7**.

Friguet, C. (2010). *Impact de La Dépendance Dans Les Procédures de Tests Multiples En Grande Dimension*. Ph.D. thesis, Rennes, AGROCAMPUS-OUEST.

Gaujoux, R. and Seoighe, C. (2010). A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics*, **11**, 367.

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., and Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**(5439), 531–537.

Grün, D., Lyubimova, A., Kester, L., Wiebrands, K., Basak, O., Sasaki, N., Clevers, H., and van Oudenaarden, A. (2015). Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, **525**(7568), 251.

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

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

Krijthe, J. H. (2015). *Rtsne: T-Distributed Stochastic Neighbor Embedding using Barnes-Hut Implementation*. R package version 0.13.

Landgraf, A. J. and 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. and 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., and 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.

Lun, A. and Risso, D. (2019). *SingleCellExperiment: S4 Classes for Single Cell Data*. R package version 1.4.1.

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

Minka, T. (2000). Estimating a Dirichlet distribution. Technical report, MIT.

Mitchell, T. J. and 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., and 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. and Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, **1**(2), 118–122.

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine Series 6*, **2**(11), 559–572.

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

Pierson, E. and 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.

Riggs, J. D. and Lalonde, T. L. (2017). *Handbook for Applied Modeling: Non-Gaussian and Correlated Data*. Cambridge University Press.

Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S., and Vert, J.-P. (2017). ZINB-WaVE: A general and flexible method for signal extraction from single-cell RNA-seq data. *bioRxiv*, page 125112.

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

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

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., and 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*.

Soneson, C. and Robinson, M. D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. *Nature Methods*, **15**(4), 255–261.

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

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and 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. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**(Nov), 2579–2605.

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

Witten, D. M., Tibshirani, R., and 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.

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

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

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

Supplementary materials

S.1 Generalization of explained variance

In the Gaussian framework, we assume that $X_{ij} \sim \mathcal{N}(\mu_{ij}, 1)$ since data are preliminary centered and scaled in PCA. Under the assumptions of independence between observations, the log-likelihood is then in matrix notation:

$$\begin{aligned}\log p(\mathbf{X} \,|\, \mathbf{M}) &= \sum_{i=1}^{n}\sum_{j=1}^{m} \log p(x_{ij} \,|\, \mu_{ij}) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} (x_{ij} - \mu_{ij})^2 \\ &= \|\mathbf{X} - \mathbf{M}\|_F^2\end{aligned}$$

where $\mathbf{M} = [\mu_{ij}]$ is the matrix of Gaussian expectation and $\|\cdot\|_F^2$ is the squared Frobenius norm. In the generalized PCA framework (Collins *et al.*, 2001), we are looking for $\mathbf{U} \in \mathbb{R}^{n\times K}$ and $\mathbf{V} \in \mathbb{R}^{m\times K}$ such that $\mathbf{M} = \mathbf{U}\mathbf{V}^T$. Thanks to Eckart and Young (1936) theorem, best $\mathbf{U}$ and $\mathbf{V}$ minimizing the Frobenius norm between $\mathbf{X}$ and $\mathbf{U}\mathbf{V}^T$ are given by Singular Value Decomposition (SVD) of $\mathbf{X}$, and optimal $\mathbf{U}$ exactly corresponds to the principal components from the PCA, which highlights the link between PCA, SVD and Gaussian framework.

In this Gaussian framework, the explained deviance defined in Equation (3) can be rewritten as

$$\%\text{dev} = \frac{\log p(\mathbf{X} \,|\, \mathbf{M} = \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T) - \log p(\mathbf{X} \,|\, \mathbf{M} = \mathbf{1}_n\bar{\mathbf{X}})}{\log p(\mathbf{X} \,|\, \mathbf{M} = \mathbf{X}) - \log p(\mathbf{X} \,|\, \mathbf{M} = \mathbf{1}_n\bar{\mathbf{X}})},$$

since the saturated model corresponds to $\mathbf{M} = \mathbf{X}$ in this case. It follows that

$$\%\text{dev} = \frac{\|\mathbf{X} - \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T\|_F^2 - \|\mathbf{X} - \mathbf{1}_n\bar{\mathbf{X}}\|_F^2}{\|\mathbf{X} - \mathbf{X}\|_F^2 - \|\mathbf{X} - \mathbf{1}_n\bar{\mathbf{X}}\|_F^2}.$$

In addition, we have that $\bar{\mathbf{X}} = 0$ thanks to the pre-centering, and the formulation becomes:

$$\begin{aligned}\%\text{dev} &= 1 - \frac{\|\mathbf{X} - \widehat{\mathbf{U}}\widehat{\mathbf{V}}^T\|_F^2}{\|\mathbf{X}\|_F^2}, \\ &= 1 - \frac{\sum_{k=K+1}^{\text{rk}(\mathbf{X})} \sigma_k^2}{\sum_{\ell=1}^{\text{rk}(\mathbf{X})} \sigma_\ell^2} \\ &= \frac{\sum_{k=1}^{K} \sigma_k^2}{\sum_{\ell=1}^{\text{rk}(\mathbf{X})} \sigma_\ell^2},\end{aligned}$$

where $\mathrm{rk}(\mathbf{X})$ is the rank of $\mathbf{X}$ and $\sigma_1 > \ldots > \sigma_{\mathrm{rk}(\mathbf{X})}$ the singular values of $\mathbf{X}$ (given by the SVD). The criterion corresponds exactly to the percentage of explained variance computed in PCA. Thus our percentage of explained deviance can be viewed as a generalization of this criterion to other distributions in the exponential family.

S.2 Identifiability issues

S.2.1 Factor order

Gamma-Poisson factor model suffers from an identifiability issue regarding the order of factors. Unlike PCA, the components of model-based factor models are not orthogonal and can not be ordered naturally since the associated likelihood is identifiable up to a permutation of factors. 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 should be all computed for every choice of dimension K, contrary to PCA.

sub

S.3 Scaling effect in GaP factor model

As stated in Dikmen and 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 using the scaling property of the Gamma distribution: if $U_{ik} \sim \mathrm{Gamma}(\alpha_{k,1}, \alpha_{k,2})$ then $\eta_k\, U_{ik} \sim \Gamma(\alpha_{k,1}, \eta_k^{-1}\alpha_{k,2})$. We show (c.f. below) that the joint log-likelihood regarding $\mathbf{U}\mathbf{H}^{-1}$ and $\mathbf{V}\mathbf{H}$ with $\mathbf{H} = \mathrm{diag}(\eta_k)_{k=1:K}$ verifies:

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

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 and Févotte, 2012). When considering zero-inflation or sparsity in the model, Equation (S.1) holds regarding the parameters of the

Gamma prior distributions and we have to consider the same precaution. However, in practice we did not encounter such sequence of diverging parameters.

Proof of Equation (S.1). We set, $\alpha^*_{k,2} = \eta_k\,\alpha_{k,2}$ and $\beta^*_{k,2} = (\eta_k)^{-1}\,\beta_{k,2}$ for fixed values $\eta_k > 0$. We use the scaling property of the Gamma distribution: if $U_{ik} \sim \text{Gamma}(\alpha_{k,1}, \alpha_{k,2})$ then $\eta_k\,U_{ik} \sim \Gamma(\alpha_{k,1}, (\eta_k)^{-1}\alpha_{k,2})$. The joint log-likelihood regarding $\mathbf{U}\mathbf{H}^{-1}$ and $\mathbf{V}\mathbf{H}$ with $\mathbf{H} = \text{diag}(\eta_k)_{k=1:K}$ is then:

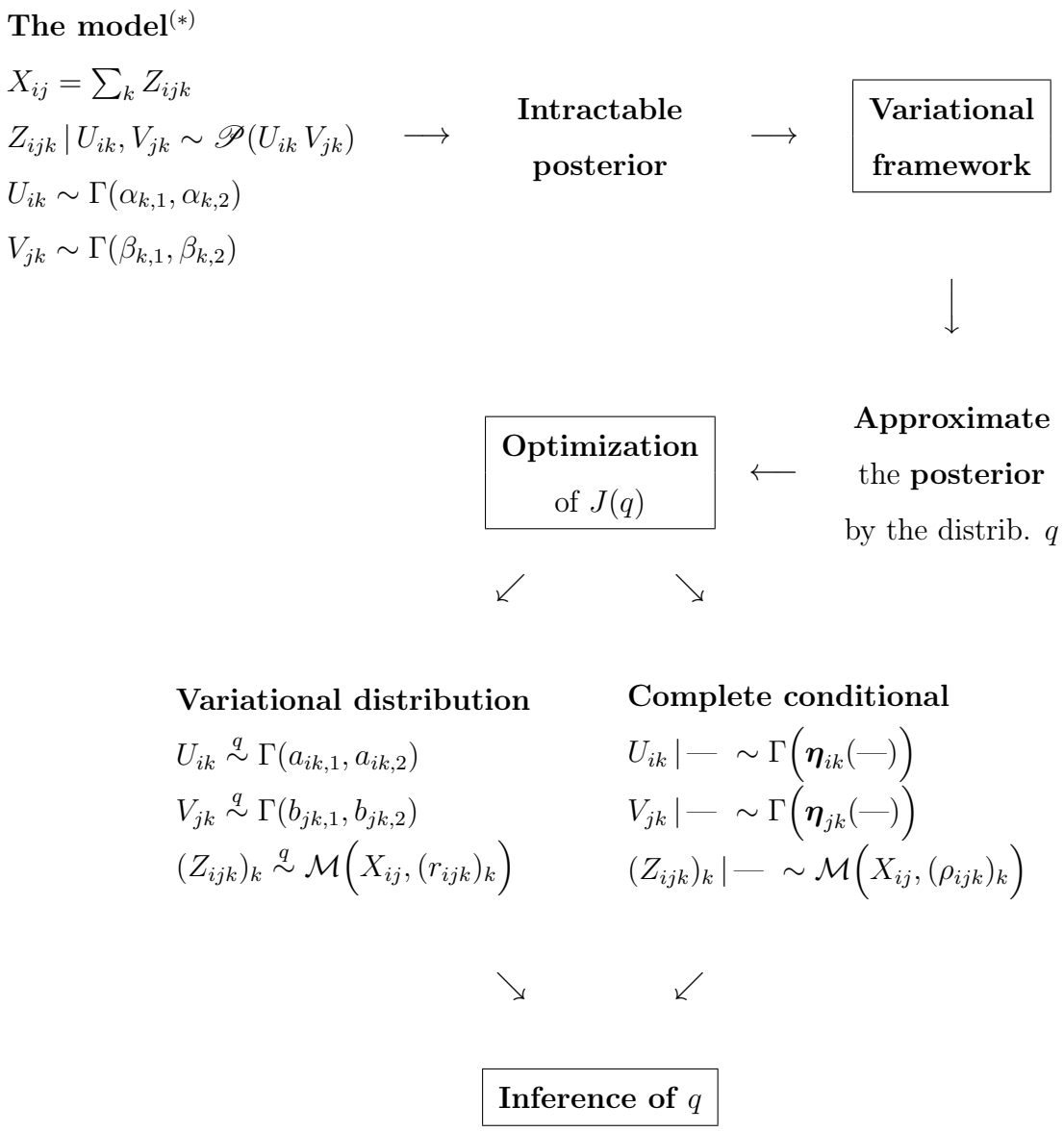
$$
\begin{aligned}
&\log p(\mathbf{X}, \mathbf{U}\mathbf{H}^{-1}, \mathbf{V}\mathbf{H} \,|\, \boldsymbol{\alpha}_1, \mathbf{H}\boldsymbol{\alpha}_2, \boldsymbol{\beta}_1, \mathbf{H}^{-1}\boldsymbol{\beta}_2) \\
&= \sum_{i,j,k} \log p\big(x_{ij} \,|\, \{(\eta_k)^{-1}\,u_{ik}, \eta_k\,v_{jk}\}_{k=1:K}\big) \\
&\quad + \sum_{i,k} \log p\big(\eta_k^{-1}\,u_{ik}\,;\,\alpha_{k,1}, \eta_k\,\alpha_{k,2}\big) \\
&\quad + \sum_{j,k} \log p\big(\eta_k\,v_{jk}\,;\,\beta_{k,1}, (\eta_k)^{-1}\,\beta_{k,2}\big) \\
&= \log p(\mathbf{X} \,|\, \mathbf{U}, \mathbf{V}) + \log p(\mathbf{U}\,;\,\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2) + \sum_{i=1}^{n}\sum_{k=1}^{K} \log(\eta_k) \\
&\quad + \log p(\mathbf{V}\,;\,\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + \sum_{j=1}^{p}\sum_{k=1}^{K} -\log(\eta_k) \\
&= \log p(\mathbf{X}, \mathbf{U}, \mathbf{V} \,|\, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2) + (n-p)\sum_{k} \log(\eta_k)
\end{aligned}
$$

S.4 Variational inference algorithm

Figure S.1 describes the variational framework (for the GaP factor model) that we extended to develop our approach.

S.4.1 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 *et al.*, 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



(*) with conditional independence between the Z_{ijk}'s and independence between the U_{ik}'s and V_{jk}'s

Figure S.1: Variational inference to approximate the posterior of the model, based on the optimization of the ELBO that required to derive the full conditional. The notation $\overset{q}{\sim}$ refers to the variational distribution.

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:

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

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:

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

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:

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

S.4.2 Derivation of variational parameters

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 $J(q)$):

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

which generalizes formulations from standard GaP factor model (Cemgil, 2009). As for variable Z_{ijk}, 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}, through 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 $\widetilde{S}_{jk} = \mathbf{1}_{\{p^S_{jk} > \tau\}}$, 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:

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

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:

$$\begin{aligned}\text{logit}(p^S_{jk}) = \text{logit}(\pi^S_j) - \textstyle\sum_i \widehat{D}_{ij}\, \widehat{U}_{ik}\, \widehat{V'}_{jk} \\ + \widehat{D}_{ij}\, \widehat{Z}_{ijk} \left(\widehat{\log U}_{ik} + \widehat{\log V'}_{jk}\right).\end{aligned}$$

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

S.4.3 Derivation of prior parameters

The hyper-parameters of priror distribution regarding U_{ik}, V_{jk}, D_{ij} and S_{jk} are updated within the M-step such that (respectively):

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

where ψ is the digamma function, i.e. the derivative of the log-Gamma function. Its inverse is computed thanks to the method proposed in Minka (2000, appendix C). 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).

S.4.4 Algorithm

Our pCMF algorithm is summarized in Algorithm S.1. In the initialization step, each variational Gamma shape parameter $a_{ik,1}$ and $b_{jk,2}$ are sampled from a Gamma distribution (Zhou *et al.*, 2012). Each variational Gamma rate parameter $a_{ik,2}$ and $b_{jk,2}$ are set to 1 (to avoid scaling effect between shape and rate parameters). Each variational dropout probability p_{ij}^D is initialized with the corresponding indicator $\delta_0(X_{ij})$. Each variational sparsity probability p_{jk}^S is initialized with the corresponding with the threshold value $\tau \in (0, 1)$. In addition, all prior hyper-parameters are initialized following update rules based on variational parameters (defined in Section 3.4 in the paper).

The convergence is assessed by computing the normalized gap between two successive parameter values across iterations. When the updates does not modify the values of the parameters, we can consider that we reach a fixed point and thus the optimum. In addition, to overcome potential issue related to local optimum, the algorithm is run several times with different random initialization and the best seed (regarding the ELBO criterion) is kept.

Algorithm S.1: Variational EM algorithm

Data: count matrix $\mathbf{X}$
Result: factors $\widehat{\mathbf{U}}$ and $\widehat{\mathbf{V}}$
Initialization: random initialization of variational parameters, prior hyper-parameters are updated accordingly
while *No convergence* **do**
  Update variational parameters (see Section 3.3 in the paper);
  Update prior pararmeters (see Section 3.4 in the paper);
end

S.5 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 $\Gamma(1, 1/\alpha_g)$ with a rate parameter depending on $\alpha_g > 0$ (different for each group). All entries U_{ik} that are not in the diagonal blocks of $\mathbf{U}$ are drawn from a Gamma distribution $\Gamma(1, 1/((1-\theta_u)\bar{\alpha})$ where $\bar{\alpha}$ is the average of the α_g's across g, and $\theta_u \in (0, 1)$ quantifies how much the groups of individuals are distinct. 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 practice, we fix $\alpha_g \in \{100, 250\}$, we use $\theta = 0.5$ or 0.8 (for low or high separation respectively) and $N = 3$ 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 m_0 and $m - m_0$ (with $m_0 \leq m$). The m_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 $\mathbf{V}^\varnothing$:

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

The ratio m_0/m sets the expected degree of sparsity in the model. In practice, we generate m_0/m from a Beta distribution, so that in average $m_0/m \in \{0.2, 0.4, 0.6, 0.8\}$ corresponding to different proportions of noisy genes (between 20 and 80% 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 M balanced groups, denoted by $\mathcal{V}_1, \ldots, \mathcal{V}_M$. We generate the corresponding matrix $\mathbf{V}^0$ with blocks on the diagonal. Each block, denoted by $\mathcal{B}_{\mathbf{V},g}$ contains m_0/M rows and K/M columns: Each entry V_{jk} in each block $\mathcal{B}_{\mathbf{V},g}$ ($g = 1, \ldots, M$) is drawn from a Gamma distribution $\Gamma(1, 1/\beta)$ with a rate parameter depending on $\beta > 0$. All entries V_{jk} that are not in the blocks on diagonal are drawn from a Gamma distribution $\Gamma(1, 1/((1-\theta_v)\beta)$, where $\theta_v \in (0, 1)$ quantifies how much the groups of individuals are distinct. Hence, each groups of individuals $\mathcal{V}_g$ corresponds to a block $\mathcal{B}_{\mathbf{V},g}$. Again, this generation pattern requires that $K > M$. In practice, we fix $\beta = 80$, we use $\theta_v = 0.8$ and $M = 2$ groups of variables.

In addition, all V_{jk} in $\mathbf{V}^\varnothing$ (noisy genes) are drawn from a Gamma distribution $\Gamma(1, 1/((1-\theta_v)\beta)$, so that $\mathbb{E}[V_{jk}]$ will not be structured according to groups.

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:

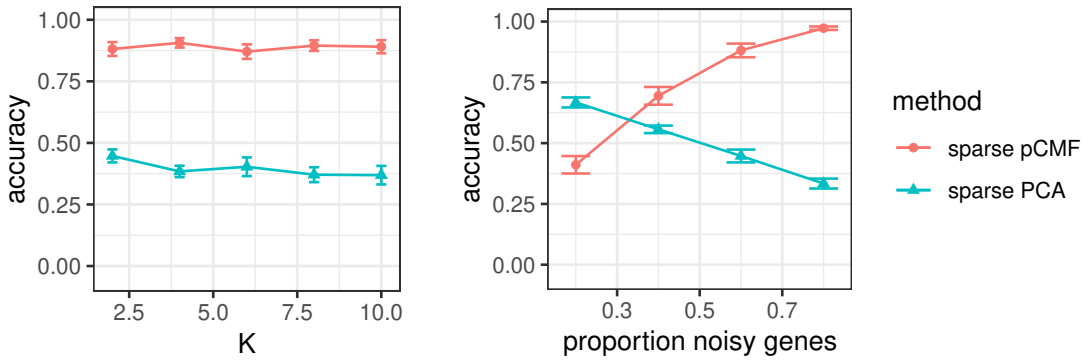
$$X_{ij} \,|\, (U_{ik}, V_{jk})_k, D_{ij} \sim (1 - D_{ij}) \times \delta_0 + D_{ij} \times \mathscr{P}(\textstyle\sum_k U_{ik}\, V_{jk})\,,$$

where the dropout indicator D_{ij} is drawn from a Bernoulli distribution $\mathcal{B}(\pi_j^D)$, 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. $\pi_j^D = 1$ for any j.

In practice, we fix $K = 40$, $n = 100$ and $m = 800$ to simulate our data. We generate different level of zero-inflation: π_j^D is drawn from a beta distribution so that in average it lies in $\{0.3, 0.5, 0.7, 0.9\}$.

S.6 Softwares

The Poisson-NMF is from the `NMF` R-package (Gaujoux and Seoighe, 2010), ZIFA from the `ZIFA` Python-package (Pierson and Yau, 2015), the sparse PCA from the `PMA` R-package (Witten *et al.*, 2009) and t-SNE from the `Rtsne` R-package (Krijthe, 2015). Computation of adjusted Rand Index was done thanks to the `mclust` R-package (Fraley and Raftery, 2002).



(a) (b)

Figure S.2: Selection accuracy depending on the dimension K (S.2a) with a proportion of noisy genes set to 60% and the proportion of noisy genes (S.2b) with K set to 2. Average values and deviation are estimated across 50 repetitions.

S.7 Additional results

S.7.1 Selection accuracy

See Figure S.2.

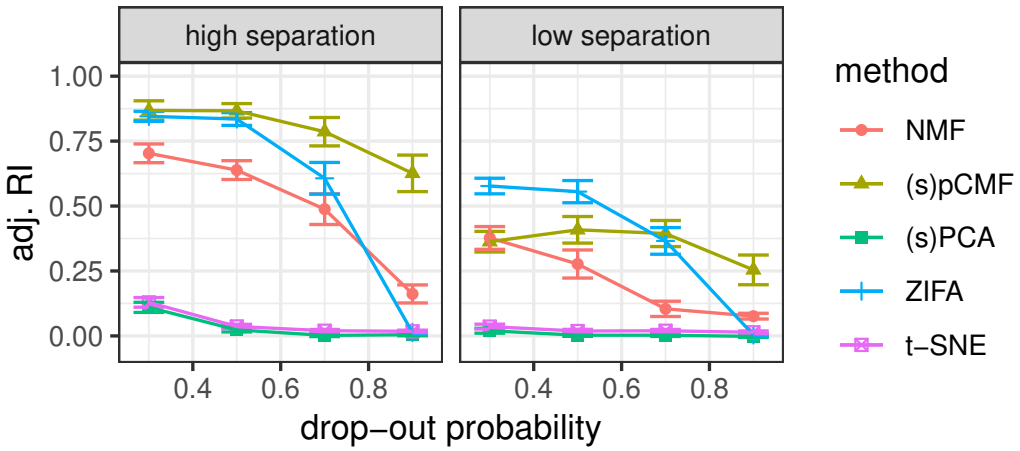
S.7.2 Clustering

See Figures S.3 and S.4. We present results on simulations with different degree of separation between the groups of individuals.

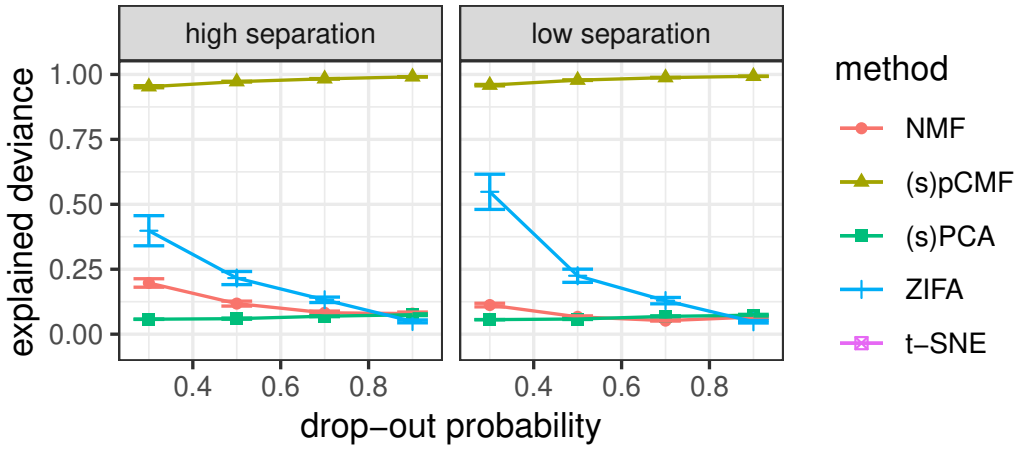
S.7.3 Computation time

Figure S.5 shows average computation time for the different methods (pCMF, Poisson-NMF, SPCA, ZIFA) for a single run on a 8-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.

Our method sparse pCMF shows comparable computation time as state-of-the-art approaches as Poisson-NMF. The npn-sparse version pCMF is slower but still faster than ZIFA and sparse PCA (because the latter requires a cross-validation step to tune a

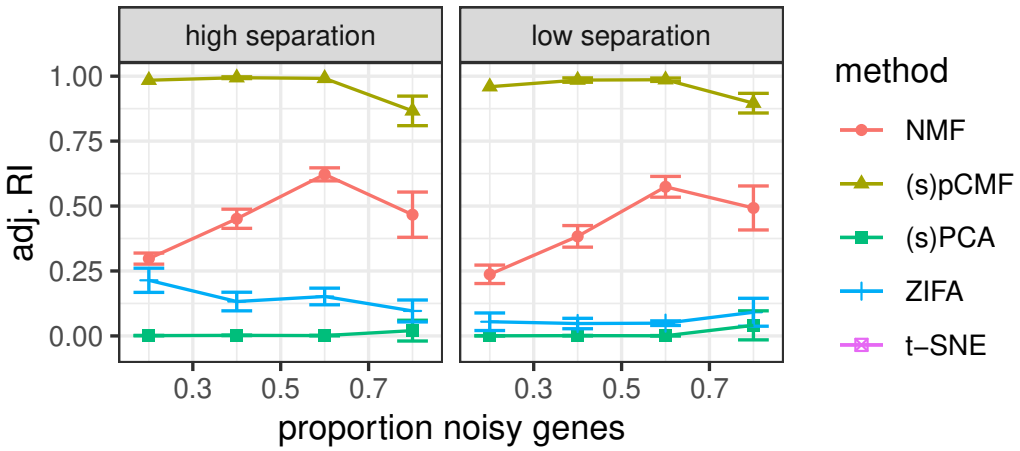


(a)

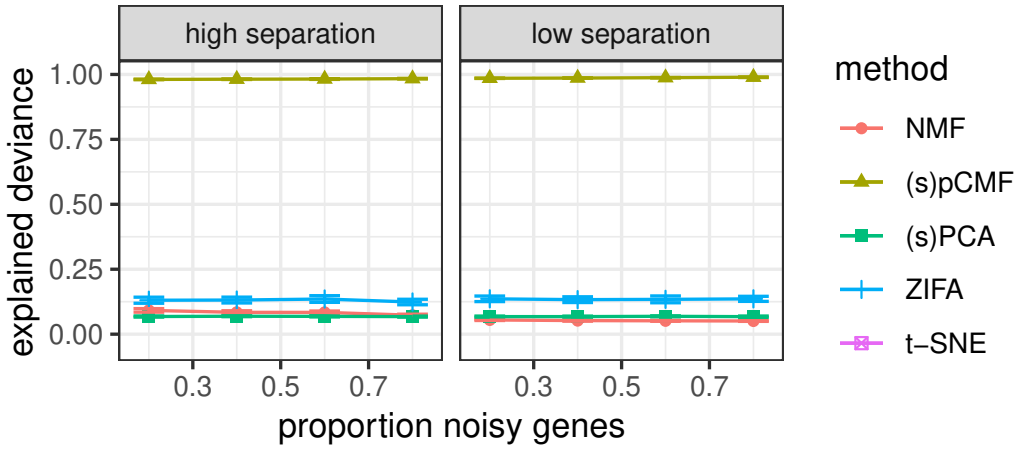


(b)

Figure S.3: Adjusted Rand Index (S.3a) for the clustering on $\widehat{\mathbf{U}}$ versus the true groups of cells; and explained deviance (S.3b) depending on the probability used to generate dropout events, for different levels of separability between cell groups. Average values and deviation are estimated across 50 repetitions.



(a)



(b)

Figure S.4: Adjusted Rand Index (S.4a) for the clustering on $\widehat{\mathbf{V}}$ versus the true groups of genes; and explained deviance (S.4b) depending on the proportion of noisy genes, for different levels of separability between cell groups. Average values and deviation are estimated across 50 repetitions.

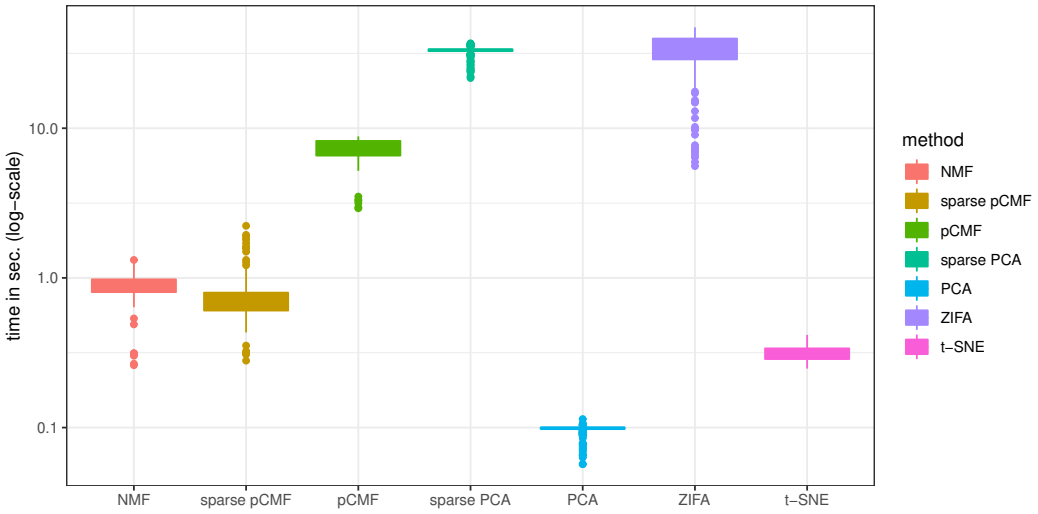


Figure S.5: Computation time on 8-CPU core for the different approaches, running time required to analyse simulated data with $n = 100$ individuals and $m = 800$ cells (50 repetitions).

penalty parameter). t-SNE is slightly faster but requires numerous run with different values for the perplexity parameter (here the timing corresponds to a run for a single perplexity value). The PCA is the gold standard regarding running time thanks to the efficiency of its algorithm based on the Singular Value Decomposition (SVD) algorithm. Packages from where the different methods can be found are detailed in Section S.6.

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

S.7.4 Standard GaP versus our ZI sparse GaP factor model

Figure S.6 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.

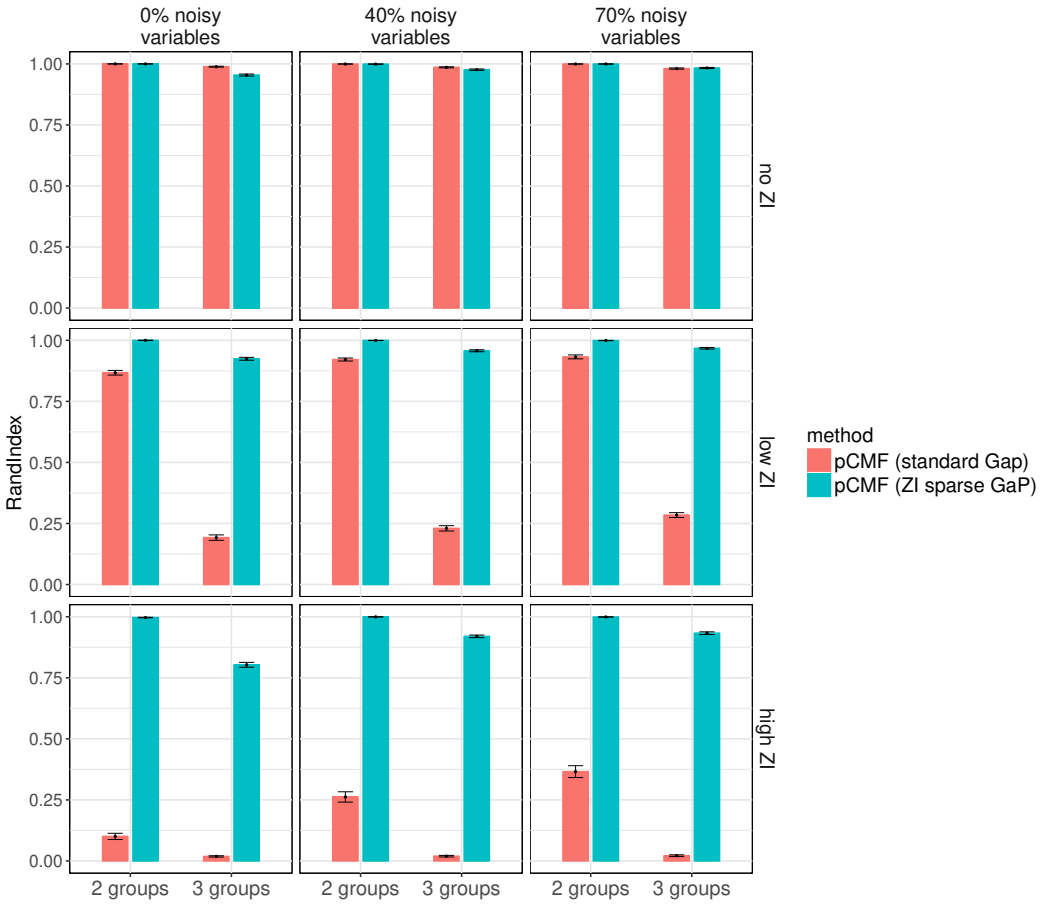


Figure S.6: Adjusted Rand Index comparing clusters found by a κ-means algorithm (applied to $\widehat{\mathbf{U}}$ with $\kappa = 2$) and the original groups of individuals, depending on the number of individual groups in the data, for different levels of zero-inflation and different proportion of noisy variables in the data. The number of components is set to $K = 10$. Data are generated with $n = 100$, $m = 1000$. Average values and deviation are estimated across 100 repetitions.

S.7.5 Additional scRNA-seq data analyses

The dataset from Baron *et al.* (2016) is available here[4]. The goldstandard and silverstandard datasets used in Freytag *et al.* (2018) can be found here[5] (we used the silverstandard dataset 5 which was the largest). The 3 previous datasets are stored based on the `SingleCellExperiment` R package Lun and Risso (2019). The dataset from Llorens-Bobadilla *et al.* (2015) was available as supplementary data of their paper. They kindly shared with us the information about cell tags.

S.7.5.1 Llorens-Bobadilla *et al.* (2015)

We illustrate the performance of pCMF on a publicly available scRNA-seq datasets of neuronal stem cells (Llorens-Bobadilla *et al.*, 2015). Neural stem cells (NSC) 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 NSC 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). The authors used a PCA approach to reveal a continuum of "activation state", from astrocytes (low activation) to amplifying progenitor cells (TAP).

As stated before, we confront pCMF with other state-of-the-art approaches. The first visual result (c.f. Figure S.9) is that pCMF provides a slightly better representation of the continuum of activation described by Llorens-Bobadilla *et al.* than PCA and t-SNE, which probably reflects a better modeling of the biological variations that exist between activation states. In practice, t-SNE was not able to highlight the different clusters of cells. 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. These qualitative results are confirmed by clustering quantitative results (c.f. Table 1 in the manuscript). The adjusted Rand Index computed after pCMF and ZIFA are similar and better than PCA.

S.7.5.2 Representation of cells

See Figures S.7 to S.9.

[4] https://hemberg-lab.github.io/scRNA.seq.datasets/mouse/pancreas/
[5] https://github.com/bahlolab/cluster_benchmark_data

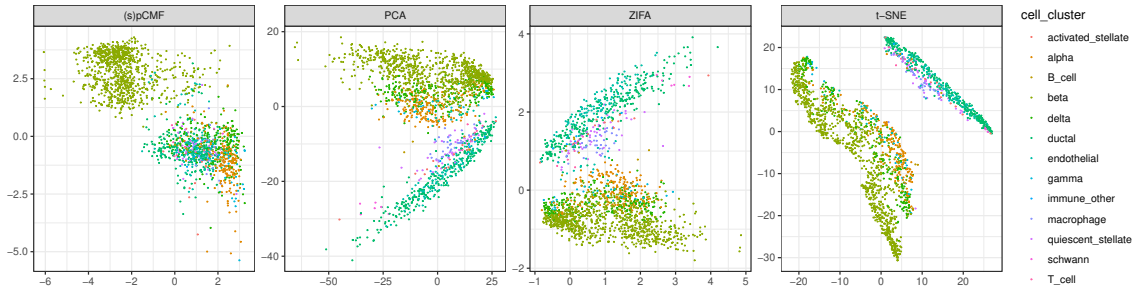


Figure S.7: Analysis of the scRNA-seq dataset from Baron *et al.* (2016), 1186 cells, 6080 genes. Visualization of the cells in a latent space of dimension 2.

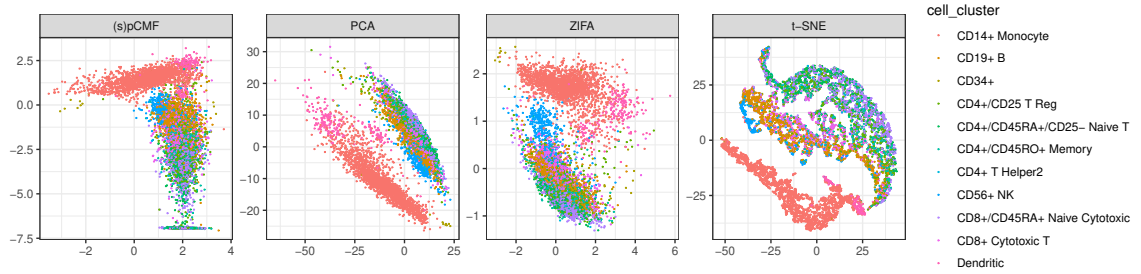


Figure S.8: Analysis of the silverstandard 5 scRNA-seq dataset from Freytag *et al.* (2018), 8352 cells, 4547 genes. Visualization of the cells in a latent space of dimension 2.

S.7.5.3 Representation of genes

See Figures S.10 to S.13.

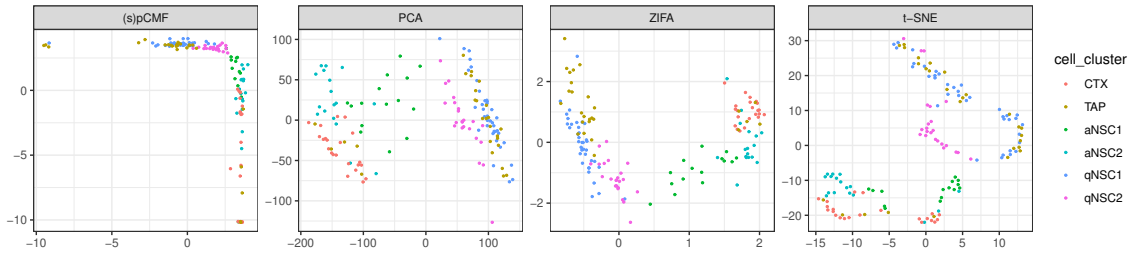


Figure S.9: Analysis of the scRNA-seq dataset from Llorens-Bobadilla *et al.* (2015), 141 cells, 13826 genes. Visualization of the genes in a latent space of dimension 2.

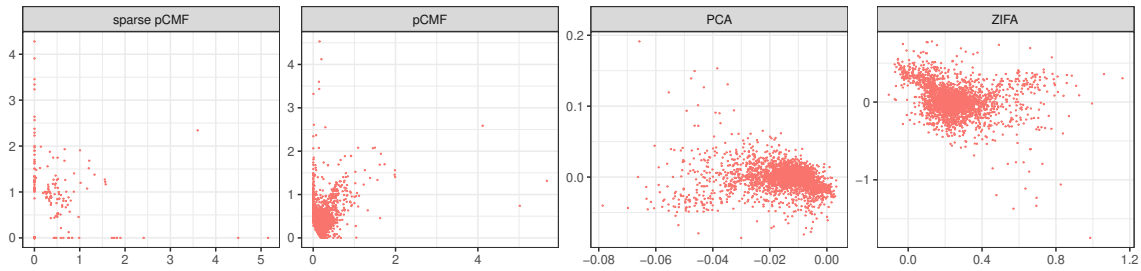


Figure S.10: Analysis of the scRNA-seq dataset from Baron *et al.* (2016), 1186 cells, 6080 genes. Visualization of the genes in a latent space of dimension 2.

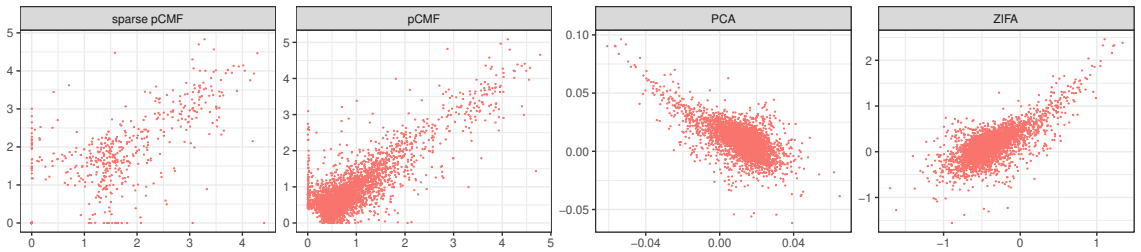


Figure S.11: Analysis of the goldstandard scRNA-seq data from Freytag *et al.* (2018), 925 cells, 8580 genes. Visualization of the genes in a latent space of dimension 2.

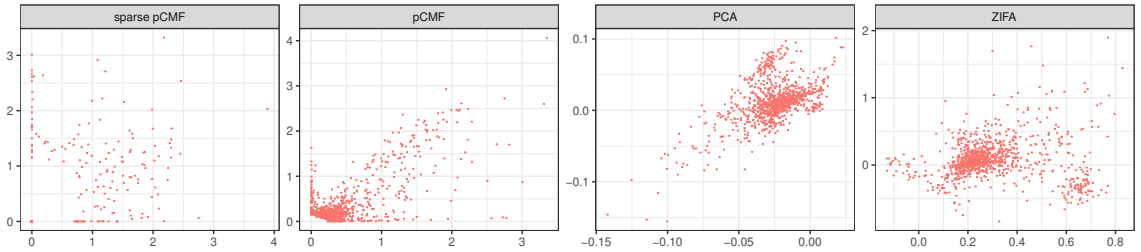


Figure S.12: Analysis of the silverstandard 5 scRNA-seq dataset from Freytag *et al.* (2018), 8352 cells, 4547 genes. Visualization of the genes in a latent space of dimension 2.

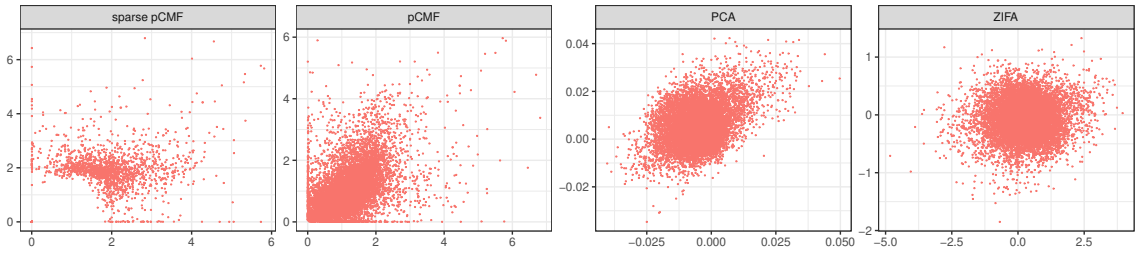


Figure S.13: Analysis of the scRNA-seq dataset from Llorens-Bobadilla *et al.* (2015), 141 cells, 13826 genes. Visualization of the genes in a latent space of dimension 2.