# Multiway latent block model for pca tensor decomposition

R.Priam[*]

November 25, 2024

## Abstract

Several latent block models (LBM) with multidimensional latent vectors for modeling blocks are discussed and BEM-like algorithms are proposed for the parameters inference from multiway matrices. The resulting extended LBM leads to summarize high-dimensional tensors for reduction and visualization in a generic multidimensional setting. These new models are for the data analysis and the visual exploration of high-dimensional big tensors.

## 1 Introduction

When data matrices are large, a clustering [1] can give a quicker access to the data contents than a method for reducing the dimensionality of the features. Combining clustering and reduction for the projection of the clusters rather than the projection of the rows or the columns is therefore an interesting requirement for data analysis. One way to fulfil this aim is to show the clusters - plus the rows or columns- on a map after clustering the data by an ad hoc algorithm and reducing the dimensions of the matrix, both separately. The family of co-clustering methods makes possible to reveal the hidden association between the rows and the columns of a data table. A simultaneous partitioning treats symmetrically the table on the contrary to a clustering of just one dimension. Algorithms were introduced earlier in the literature in [2] and [3, 4]. There is also the information-theoretic co-clustering method [5] and its generalization [6]. Other approaches are for instance the general method for prediction [7] or the non negative matricial decompositions [8, 9, 10]. In [11, 12], a latent block mixture model and its inference by a variational generalized EM algorithm are introduced. Re-parameterization of these models via latent variables [13] bring new properties to these models for: rows and columns visualizations, dimensions reductions or even factorizations as proposed herein.

### Foundation Models

For data reduction, principal component analysis (PCA) [14] is intensively studied. Robust versions of this method are developped in order to get rid of the influencial outliers but a skew version is also appealing for non elliptical distribution. Robustness in factorization is introduced via a specific modeling of the noise with hierarchical Laplace models in order to get better estimations of the parameters. Factorization [15] is a living activity for reducing the dimensionality of matrices. For helping at denoising and clustering, principal component analysis aims at the construction of synthetic variables from the original variables or columns of the data matrix. These new variables maximize an explained variance via a projection from the full data cloud to a subspace spanning the columns space. The limited number of new variables or components in the new vectors that replace the (possiby high-dimensional) original one leads to the property of reducing the dimensionality. A new matrix is available with less columns (or rows) than the original matrix. For all these reasons, this method is usefull is numerous research problems from an applied statistical analysis to a pre-reduction for an other method such that clustering. More formally, let $\mathbf{x}$ the numerical matrix with $n$ rows $\mathbf{x}_i$ and cells $x_{ij}$. PCA has several definitions actually in the literature according to each purpose. There can be bayesian priors for the reduced vectors, orthogonality for the loading matrix, or a factorization without other hypothesis. The following definitions are widely considered:

- For centered data, PCA is a matricial method [16] which is originally defined as solving for:

  $$\begin{array}{rcll} R^* & = & argmin_R & \sum_i \|\mathbf{x}_i - RR^T\mathbf{x}_i\|^2 \\ & = & argmin_R & \|\mathbf{x} - RR^T\mathbf{x}\|^2 \,. \end{array} \tag{1}$$

  This is equivalent to define as minimizing the reconstruction error of the projected data points into the new subspace, a linear transformationfor reducing the space of the columns. Eventually, orthogonality supposes $R^TR = \mathbf{I}$ which often improves the results.

- A factorization [15] is written as follows,

  $$\begin{array}{rcl} (\hat{\mathbf{u}}, \hat{\mathbf{v}}) & = & argmin_{\mathbf{u},\mathbf{v}} ||\mathbf{x} - \mathbf{u}^T\mathbf{v}||_F^2 \\ & = & argmin_{\mathbf{u},\mathbf{v}} \sum_{i,j} (x_{ij} - u_i^T v_j)^2 \,. \end{array}$$

  A solution is found by regression for solving for $\mathbf{u}$ and keeping $\mathbf{v}$ constant, and similirly for $\mathbf{v}$ when $\mathbf{u}$ is constant.

- In Probabilistic PCA (PPCA [17], it may be written also for non centered data:

  $$(\hat{\mu}, \hat{W}) \quad = \quad argmin_{\mu,W} \sum_i \|\mathbf{x}_i - \mu - \mathbf{W}\mathbf{y}_i\|^2 \,.$$

  This is the approach used in the probabilistic setting of PCA where this notation is more usual in this family of models.

[*] rpriam@gmail.com.

Note that it may not be required the hypothesis of orthogonality anymore for some authors. Without this constraint, The matrix $\mathbf{W}$ is not orthogonal hence the latent variables $\mathbf{y}_i$ are not corresponding to the solution of PCA, like in an usual autoencoder [18] with deep learning.

### From robust PCA to factorization LBM

As the usual PCA is based on a Gaussian hypothesis coming from the Euclidean distance involved, it is known that this distribution limits the resistance against the influence of the outliers. These later ones are not well modeled: they pollute the parameters inference and leads to erroneous values. Two distributions - t-Student and usual Laplace distributions - have often been considered previously in the literature [19, 20, 21], with sometimes also their skewer alternatives. Skewness can be modeled for even more robustness for non elliptical error shapes and quantile are able to lead to several solutions in complement to the median one. To our knowledge, for skew PCA only the model in [22] has been proposed while for a quantile PCA never a solution has been proposed in the past. Many robustification are possible for principal component analysis [23, 24]. In particular, following RobRSVD [25] (see also R package with same name at CRAN) for singular value decomposition, a weigthing scheme is obtained when minimizing an objective function where the Euclidean distance is replaced with a more robust distance. This robust approach leads to introduce weigths, $\rho_{ij} = \frac{\rho'(x_{ij}-u_i^T v_j)}{x_{ij}-u_i^T v_j}$ associated to the usual Euclidean distance. The difference is mainly to not handle the quantiles but only the robustification. Numerous alternative functions have been proposed in the literature.

For all these reasons, when a reduction is addressed, robustness is of first interest when looking for reduced latent vectors (see [13]) in the latent block models, as considered herein.

# 2 Multiway Latent Block Models

A review of the latent block model and its general definition summarizes briefly the modeling foundation before presenting our proposal in this section.

## 2.1 Former model and algorithm

### Model

Within the context of the classical mixture model, a partition of $I$ into $g$ clusters is represented by the binary classification matrix $\mathbf{z} = (z_{ik})_{n\times g}$ such that $\sum_{k=1}^{g} z_{ik} = 1$ and $z_{ik} = 1$ indicates the component of the row $i$. Just as $I$ is partitioned into $g$ clusters, columns can be partitioned into $m$ clusters by the binary classification matrix $\mathbf{w} = (w_{j\ell})_{d\times m}$. If the most usual clustering methods deal with clustering of only the set $I$ or eventually $J$, co-clustering is interested in the clustering of both. The $n \times d$ random variables generating the observed $x_{ij}$ cells of the data matrix are assumed to be independent in LBM, once $\mathbf{z}$ and $\mathbf{w}$ are fixed. The set of all possible assignments $\mathbf{w}$ of $J$ (resp. $\mathbf{z}$ of $I$) is denoted $\mathcal{W}$ (resp. $\mathcal{Z}$). Here the data $\mathbf{x}$ is set of cells $(x_{11}, x_{12}, \ldots, x_{ij}, \ldots, x_{nd})$, such as these cells are independent conditionnaly to the row and columns labels. The two sets of possible assignments $\mathbf{w}$ and $\mathbf{z}$ aggregate the cells of the matrix $\mathbf{x}$ into a number of contiguous, non-overlapping blocks. The following decomposition is obtained [11] by independence of $\mathbf{z}$ and $\mathbf{w}$, by summing over all the assignments $\mathcal{Z} \times \mathcal{W}$:

$$f_{LBM}(\mathbf{x};\theta) = \sum_{(\mathbf{z},\mathbf{w})} \prod_{i,k} p_k^{z_{ik}} \prod_{j,\ell} q_\ell^{w_{j\ell}} \prod_{i,j,k,\ell} \varphi(x_{ij};\alpha_{k\ell})^{z_{ik}w_{j\ell}}\,.$$

Here $\varphi(.;\alpha_{k\ell})$ is a probability function defined on the real line $\mathbb{R}$ or the set of integers $\mathbb{N}$, while $\{\alpha_{k\ell}\}$ are unknown parameters. The vectors of the probabilities $p_k$ and $q_\ell$ that a row and a column belong to the $k^{\text{th}}$ component and to the $\ell^{\text{th}}$ component are respectively denoted $\mathbf{p} = (p_1, \ldots, p_g)$ and $\mathbf{q} = (q_1, \ldots, q_m)$. The set of parameters is denoted $\theta$ and is compound of $\mathbf{p}$ and $\mathbf{q}$ plus $\alpha$ which aggregates the $g \times m$ scalar $\alpha_{k\ell}$ via the matrix $(\alpha_{k\ell})$. Hereafter, to simplify the notation, the sums and the products relating to rows, columns or clusters will be subscripted respectively by the letters $i$, $j$, $k$, or $\ell$ for matrices and $i_1$,$i_1$,$\cdots$,$i_m$ and $k_1$,$k_1$,$\cdots$,$k_m$ for tensors, without indicating the limits of variation, which are implicit. The set of parameters $\theta$ of the model can be estimated by maximizing the log-likelihood:

$$L(\theta;\mathbf{x}) = \log f_{LBM}(\mathbf{x};\theta)\,.$$

A particular distribution is needed for the cells, for continuous, binary and contingency matrices with originally Gaussian, Bernoulli and Poisson distributions where the parameters are expectations plus sometimes other parameters, say ($\mu_{k\ell}, \sigma_{k\ell}$, $\alpha_{k\ell}$ and $\lambda_{k\ell}^{ij} = \mu_i \nu_j \alpha_{k\ell}$ respectively. Note that the effects $\mu = (\mu_1, \ldots, \mu_n)$ and $\nu = (\nu_1, \ldots, \nu_d)$ are assumed equal to the following constant margin totals by rows and by columns, $\mu_i = \sum_j x_{ij}$ for $i \in I$ and $\nu_j = \sum_i x_{ij}$ for $j \in J$ for contingency matrices.

### Objective function and optimization algorithm

For the proposed model with the introduced constraints, we aim to address the problem of the estimation of the parameters by a maximum likelihood (ML) approach such that:

$$\hat{\theta} = argmax_\theta L(\theta;\mathbf{x})\,.$$

Let us focus on the estimation of a value of $\theta$ by the maximum likelihood approach associated to the block mixture model. For this model, the complete data are $(\mathbf{x},\mathbf{z},\mathbf{w})$ where the unobservable vectors $\mathbf{z}$ and $\mathbf{w}$ are the labels. The complete log-likelihood of $(\mathbf{x},\mathbf{z},\mathbf{w})$ leads to an algorithm EM where the E-step is intractable, thus an alternative approach has been proposed in the literature.

An approach based on a Generalized EM and a variational approximation by the product $c_{ik}^{(t)} d_{j\ell}^{(t)}$ has been proposed [26] previously in the literature, and named *Block EM* (BEM). Here it is denoted the variational probabilities $c_{ik}$ such that $\sum_k c_{ik} = 1$, and also $d_{j\ell}$ such that $\sum_\ell d_{j\ell} = 1$. Their matricial representations are respectively $\mathbf{c} = (c_{ik})$, and $\mathbf{d} = (d_{j\ell})$, such that the variational distribution for the clustering is defined as $Q_{(\mathbf{c},\mathbf{d})}(\mathbf{z},\mathbf{w}) = \prod_{i,k}(c_{ik})^{z_{ik}} \prod_{j,\ell}(d_{j\ell})^{w_{j\ell}}$. Then, by the Jensen inequality a bound $\mathcal{F}(\mathbf{c},\mathbf{d};\theta)$ can be defined. The algorithm proceeds by defining a lower bound of the log-likelihood (see [26] and appendix) and repeats until convergence the two following steps:

**E-step** The posterior probabilities $\mathbf{e} = (\mathbf{c},\mathbf{d})$ are found at the current time (with the normalizing constraint to one). By maximizing $\mathcal{F}$ with respect to $c_{ik}$ and $d_{j\ell}$. the resulting posterior probabilities are estimated with the dependent equations:

$$\begin{aligned} c_{ik}^{(t)} &\propto p_k^{(t)} \exp\left(\sum_{j,\ell} d_{j\ell}^{(t)} \log\varphi(x_{ij};\alpha_{k\ell}^{(t)})\right), \\ d_{j\ell}^{(t)} &\propto q_\ell^{(t)} \exp\left(\sum_{i,k} c_{ik}^{(t)} \log\varphi(x_{ij};\alpha_{k\ell}^{(t)})\right). \end{aligned}$$

Here the probabilities are hence obtained as a solution of the fixed point relations after initializing with previous values.

**M-step** A temporary value of the parameters is found at the new current time. By maximizing $\mathcal{F}$ with respect to $\Omega$, the objective function to maximize is:

$$\begin{aligned} \tilde{Q}_{LBM}(\theta,\theta^{(t)}) &= \textstyle\sum_{i,j,k,\ell} c_{ik}^{(t)} d_{j\ell}^{(t)} \log\varphi(x_{ij};\alpha_{k\ell}) \\ &\quad + \textstyle\sum_{i,k} c_{ik}^{(t)} p_k + \sum_{j,\ell} d_{j\ell}^{(t)} q_\ell . \end{aligned}$$

Here, the posterior probabilities $c_{ik}^{(t)}$ and $d_{j\ell}^{(t)}$ are available from E-step, this results into the criterion also denoted $\tilde{Q}$. For $k = 1,...,g$ and $\ell = 1,...,m$, it is denoted the aggregated statistics, $x_{k\ell}^{(t)} = \sum_{i,j} c_{ik}^{(t)} d_{j\ell}^{(t)} x_{ij}$, $\mu_k^{(t)} = \sum_i c_{ik}^{(t)} \mu_i$, $\nu_\ell^{(t)} = \sum_j d_{j\ell}^{(t)} \nu_j$. Given $\theta^{(t)}$, the quantities $c_{ik}^{(t)}$ (resp. $d_{j\ell}^{(t)}$) are the posterior probabilities that a row (resp. a column) belongs to the block $k\ell$. When solving for maximizing (2), the solution for $\alpha$ can be written $\alpha_{k\ell}^{(t)} = x_{k\ell}^{(t)}/\mu_k^{(t)}\nu_\ell^{(t)}$ in Bernoulli and Gaussian cases. The solution for the mixing coefficients is also obtained as, $p_k^{(t)} = n^{-1}\sum_i c_{ik}^{(t)}$ and $q_\ell^{(t)} = d^{-1}\sum_j d_{j\ell}^{(t)}$ respectively if they were not taken constant. Hence, the parameters are estimated by an iterative way, the BEM algorithm proceeds by an alternated maximization of $\tilde{Q}$ and converges to a final solution which maximizes (locally) the log-likelihood of the latent block model. A hat is added to each parameter or statistics which is estimated and found at the final stage of the inference algorithm.

## 2.2 Multiway version and parameterization

Co-clustering methods are dedicated to the detection of the main homogeneous sub-structures in a data cloud when the variables can be clustered. At the end, one has a small matrix $\alpha_{k\ell}$ which can be considered as a very synthetic view of the dataset. The multiway version changes a 2-dimensional problem to a $m$-dimensional problem, for clustering a tensor $x_{i_1,\cdots,i_m} = I_m$ with $m$ dimensions, say a shape $n_1 \times \cdots, n_m$. Such that when denoting $I_m = \{i_1,\cdots,i_m\}$ and $\mathcal{K}_m = \{k_1,\cdots,k_m\}$ where $k_\ell \in \{1,\cdots,g_\ell\}$ for the number of clusters $g_\ell$ of the dimension $\ell$ of the tensor, one gets for the new random variables/matrices for the clustering along of each dimension, $\mathbf{z}_\ell$ for $\ell \in \{1,\cdots,m\}$, that the following new involved expressions.

### Model

The new required expressions are for the likelihood $f_{WLBM}$, the update formula of the corresponding variational probabilities $c_{i_\ell k_\ell}$, and the maximization for the M-step optimizing w.r.t the parameters $\alpha_{\mathcal{K}_m}$,

$$f_{WLBM}(\mathbf{x};\theta) = \sum_{(\mathbf{z}_1,\cdots,\mathbf{z}_m)} \prod_{\ell=1}^{m} \prod_{z_{i_\ell,k_\ell}} p_{k_\ell}^{z_{i_\ell k_\ell}} \prod_{I_m,\mathcal{K}_m} \varphi(x_{i_1,\cdots,i_m};\alpha_{z_{k_1},\cdots,z_{k_m}}).$$

This leads also to the variational bound which is maximized w.r.t. $m$ clustering variables instead of just 2 ones before,

$$c_{i_\ell k_\ell}^{(t)} \propto p_{k_\ell}^{(t)} \exp\left(\sum_{I_m\backslash i_\ell,\mathcal{K}_m\backslash k_\ell} c_{i_o k_o}^{(t)} \log\varphi(x_{I_m};\alpha_{\mathcal{K}_m}^{(t)})\right).$$

Similarly, just the maximum may be preferred for a classifying version of BEM, called BCEM, for binary variables instead of fuzy ones. By maximizing $\mathcal{F}$ with respect to $\Omega$, the objective function becomes:

$$\begin{aligned} \tilde{Q}_{WLBM}(\theta,\theta^{(t)}) &= \textstyle\sum_{I_m,\mathcal{K}_m} c_{i_1k_1}^{(t)} \cdots c_{i_mk_m}^{(t)} \log\varphi(x_{I_m};\alpha_{\mathcal{K}_m}^{(t)}) \\ &\quad + \textstyle\sum_{\ell=1}^{m} \sum_{i_\ell,k_\ell} c_{i_\ell k_\ell}^{(t)} p_\ell . \end{aligned}$$

There was just changed the indices, as the algebra remains the same: this defines our multiway latent block model and its corresponding parameters inference. The new M-step solves for the optimization:

$$\theta^{(t+1)} = \tilde{Q}_{WLBM}(\theta,\theta^{(t)})$$

When constraints are added, multidimensional latent vectors are estimated instead of directly the centers from the blocks.

### Factorization

For the decomposition of the tensor to a dimension $r$, the more direct way is when following the classical tensor algebra [27] as follows with a function $\phi(.)$ such as indentity, exponential or sigmoidal ones,

$$\alpha_{z_{k_1},\cdots,z_{k_m}} = \phi\left(\sum_{s=1}^{r} \alpha_{k_1 s} + \cdots + \alpha_{k_m s}\right).$$

This extends the scalar product between vectors [13] (as previously proposed) to a tensor decomposition with the latent matrices, with $\alpha_{k_\ell r}$ to be estimated for all $\ell$ and $s$. Next, it is presented an illustration of our proposed approach for tensor with Gaussian distributions for tensors with continuous values.

# 3 Example of factorization models

This section proposes diverse models for continuous data, with the normal cells and laplace cells[1] in order to cluster and factorize tensors. They are suitable for summarizing tensors of three dimensions, which may be more common in practice, but the parameters inference ask for intensive computations.

## 3.1 Gaussian tensor latent block factorization model

In this subsection, solutions are provided for the normal distribution, with $m = 3$. In LBM, the completed data are taken to be the vector $(\mathbf{x}, \mathbf{z}_1; \mathbf{z}_2, \mathbf{z}_3)$ where the latent variables $\mathbf{z}_1$,$\mathbf{z}_2$, and $\mathbf{z}_3$ are the random labels for the rows and the columns respectively.

### Model

The classification log-likelihood can then be written:

$$L_C(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3; \mathbf{x}, \theta) \quad = \quad \log P(\mathbf{x}|\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) + \sum_{\ell=1}^{3} \log P(\mathbf{z}_\ell)\,.$$

The different density and mass functions are:

$$\begin{aligned}
\varphi_N(u; \mu, \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} \\
P(\mathbf{z}) &= \prod_{\ell=1}^{3} \prod_{i_\ell, k_\ell} (p_{k_\ell})^{z_{i_\ell k_\ell}} \\
P(\mathbf{x}|\mathbf{z}, \mathbf{w}) &= \prod_{I_m, \mathcal{K}_m} \varphi_N(x_{i_1 i_2 i_3}; \mu_{k_1 k_2 k_3}, \sigma_{k_1 k_2 k_3})\,.
\end{aligned}$$

The set of parameters is denoted $\theta = (\mathbf{p}, \alpha)$. Here, the probabilities $p_{k_\ell}$ that a dimension of the tensor belongs to the $k_\ell^{\text{th}}$ component are aggregated in $\mathbf{p}_\ell = (p_{k_\ell 1}, \ldots, p_{k_\ell g_\ell})$, while $\alpha$ aggregates the parameters from all the p.d.f. of each block in the tensor, say $\alpha_{k_1 k_2 k_3}$. As in the Gaussian case of the non symmetric co-clustering model [28] but in three dimensions, the p.d.f. for the block $(k_1 k_2 k_3)$ is such as $\alpha_{k_1 k_2 k_3} = (\mu_{k_1 k_2 k_3}, \sigma^2_{k_1 k_2 k_3})$ with the mean $\mu_{k_1 k_2 k_3}$ and the standard-deviation $\sigma_{k_1 k_2 k_3}$. The model is called G-WLBM with aggregating tensors parameters: $\Sigma = (\sigma^2_{k_1 k_2 k_3})_{g_1 \times g_2 \times g_3}$ and $\mu = (\mu_{k_1 k_2 k_3})_{g_1 \times g_2 \times g_3}$. The probability density function (p.d.f.) of a latent block model is defined as the following decomposition. It is obtained by independence of $\mathbf{z}_1$, $\mathbf{z}_1$ and $\mathbf{z}_3$, by summing over all the assignments [11]:

$$f_{WLBM}(\mathbf{x}; \theta) = \sum_{(\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3) \in \mathcal{Z}_1 \times \mathcal{Z}_2 \times \mathcal{Z}_3} P(\mathbf{x}|\mathbf{z}_3, \mathbf{z}_3, \mathbf{z}_3) P(\mathbf{z}_1) P(\mathbf{z}_2) P(\mathbf{z}_3)\,.$$

The set of all the possible assignments is denoted $\mathcal{Z}$ for $I$ and $\mathcal{W}$ for $J$. The log-likelihood is as follows:

$$L(\mathbf{x}; \theta) = \log f_{WLBM}(\mathbf{x}; \theta).$$

The block model is dramatically more parsimonious than the usual mixture model where each dimension of the data table is modeled separately. Next paragraph, the new adapted criterion and algorithm for an estimation of the parameters are presented.

### Objective function, EM and BEM

For co-clustering, one generally addresses the problem of parameters estimation by a maximum likelihood (ML) approach such that the log-likelihood $L$ is maximized via a bound. The log-likelihood of the completed data is considered for the inference by taking benefit of the introduced latent variables for the labelling of the cells. By following the *Block EM* (BEM) as EM [29] is intractable [26], a similar variational approximation is available for a tensor. This algorithm proceeds by maximizing a surrogate objective function which lower bounds $L$ and converges towards a final solution which maximizes (locally) the log-likelihood of the latent block model. For the Gaussian case, the set of parameters is $\theta = (\alpha, \Sigma, \mathbf{p})$ with the following objective function written,

$$\begin{aligned}
\tilde{Q}(\theta|\theta^{(t)}) \quad &= \quad \textstyle\sum_{I_3, \mathcal{K}_3} c^{(t)}_{i_1 k_1} c^{(t)}_{i2 k_2} c^{(t)}_{i3 k_3} \left[ -\frac{(x_{i_1 i_2 i_3} - \mu_{k_1 k_2 k_3})^2}{2\sigma^2_{k_1 k_2 k_3}} - \frac{\log(2\pi\sigma^2_{k\ell})}{2} \right] \\
&\quad - \quad \textstyle\sum_{\ell=1}^{3} \sum_{i_\ell, k_\ell} c^{(t)}_{i_\ell k_\ell} \log c^{(t)}_{i_\ell k_\ell} + \sum_{\ell=1}^{3} \sum_{k_\ell} c^{(t)}_{k_\ell} \log p_{k_\ell}\,.
\end{aligned}$$

The solutions are in closed form such that when introducing latent variable in order to reduce the data tensor.

## 3.2 Laplace tensor latent block factorization model

The Euclidean distance is known to be non robust because an outlier has an influence which comes with the square of the norm. To solve this issue, the median or more generally, a quantile is preferred. The later one leads to several solutions instead of just the median one which is included among the possible solutions. For treating skewness or quantile within the principal component analysis framework, a hierarchical Laplace model is appealing as an expectation-maximization algorithm brings closed-form updates of the parameters for the inference.

### Laplace PCA model for matrix without clustering

This part introduces a model and algorithm for the inference for a skewed quantile principal component analysis with also the anotations and the general modeling framework. This is a baseline for the proposed new model with tensors and clustering along the dimensions.

- **Parametric distribution:** In the literature a main tool in quantile modeling is the check function, $\rho_p(r) = rI(r \geq 0) - (1-p)r$, which is difficult to optimize directly hence a Bayesian alternative is preferred in the following. A random variable $Z$ has an Asymmetric Laplace distribution with location parameter $\mu$, scale parameter $\sigma > 0$ and skweness parameter $p \in (0, 11)$, when the probability density function (p.d.f) is denoted $ALD(\mu, \sigma, p)$. For a cell sample $x_{ij}$ generated from an ALD, it is written that:

$$\varphi_L(u|\mu, \sigma, p) = \frac{p(1-p)}{p} exp\left\{ -\rho_p\left(\frac{u-\mu}{\sigma}\right)\right\}. \qquad (2)$$

[1] This part (with also the end of the introduction) is adapted from an old draft from same author on robust pca. As the formula were not double-checked or implemented they are given for information and provisory.

For the $(i,j)$th cell in order to retrieve a factorization -of the type PCA-, it is chosen:

$$\mu = \mu_j + u_i^T v_j \,,$$

The hypothesis is generative and independant identical distributions (i.i.d.) for the whole data matrix. Maximizing the corresponding likelihood is equivalent to minimize the check function which is not smooth. A scaled Gaussian model has been proposed recently via an exponential distribution in order to get a smooth objective function under a probabilistic setting.

- **Introduction to hierarchical ALD:** Let for the $(i,j)^{\text{th}}$ cell of $x$, have two variable stochastically independant,

$$\begin{array}{rcl} w_{ij} & \sim & exp(\sigma) \\ z_{ij} & \sim & N(0,1)\,. \end{array}$$

The variable $x_{ij}$ has generated a sample -with same notation $x_{ij}$- which is a cell in the data matrix. It is drawn from a variable ALD which can be modeled as follows:

$$x_{ij}|w_{ij} = \mu + \varphi_p w_{ij} + \tau_p \sqrt{\sigma w_{ij}} Z_{ij} \,.$$

Where,

$$\varphi_p = \tfrac{1-2p}{p(1-p)} \text{ and } \tau_p = \tfrac{2}{p(1-p)} \,.$$

This can be summarized into:

$$\begin{array}{rcl} x_{ij}|w_{ij} & \sim & N(\mu_{ij} + \varphi_p w_{ij}, \tau_p^2 \sigma w_{ij}) \\ w_{ij} & \sim & exp(\sigma)\,. \end{array}$$

Next, it is explained how we can take advantage the machinery of the expectation-maximization algorithm for the maximization of the likelihood via $w_{ij}$ as the missing variables. For lighter notation, the later ones are denoted in lowercase for the random variables and the observations.

The expectation for $w_{ij}^s$ are found as follows according to the literature. Let $K_\nu(.)$ denotes the modified Bessel function of the third kind, and define,

$$\delta_{ij} = \tfrac{|x_{ij} - u_i^T v_j|}{\tau_p \sqrt{\sigma}},\, \gamma = \tfrac{\tau_p}{2\sqrt{\sigma}} \text{ and } \lambda_{ij} = \delta_{ij}/\gamma.$$

From the properties of the related distribution, the expectations involved for the missing variables are denoted at $t^{\text{th}}$ step, for $s \in \{1,-1\}$ with values,

$$\begin{array}{rcl} e_{ij,s} & = & \mathbb{E}[w_{ij}^s|\theta, x] \\ & = & \left(\frac{\delta_{ij}}{\gamma}\right) \frac{K_{1/2+s}(\lambda_{ij})}{K_{1/2}(\lambda_{ij})}\,. \end{array} \tag{3}$$

This comes from the conditional distribution which follows a generalized inverse Gaussian (GIG) distributions. These expectation are usefull in the inferencial procedure.

- **Functions $\mathcal{L}$ and $\mathcal{Q}$:** With $\theta$ aggregating the parameters $\mu = (\mu_j)_{1\leq j\leq m}^T$, **u**, **v**, and $\sigma$, the likelihood to maximize is written by the product of the densities for all the observations in the samples. This leads to this following objective function:

$$\mathcal{L}(\theta) \;=\; \textstyle\prod_{ij} f(x_{ij}|\mu_j + u_i^T v_j, \sigma, p)\,. \tag{4}$$

The logarithm of the complete likelihood is written by adding the hidden variables which are not marginalized before the logarithmic transformation. This will allow to optimize a surrogate objective function coming from an expectation. The function in stake is written as:

$$\begin{array}{rl} & \mathcal{L}(\theta, \mathbf{w}|x) \\ = & \sum_{i,j} \log\left\{f(x_{ij}|w_{ij})\right\} \\ = & \sum_{i,j} \log\left\{ \frac{e^{-\frac{1}{2\tau_p^2\sigma w_{ij}}[x_{ij}-\mu_j-u_i^T v_j-\varphi_p w_{ij}]^2}}{\sqrt{2\pi\tau_p^2\sigma w_{ij}}} \frac{1}{\sigma} e^{-\frac{w_{ij}}{\sigma}} \right\} \\ = & \sum_{i,j} \left\{ -\frac{1}{2}\log(2\pi\tau_p^2\sigma^3) - \frac{1}{2}\log(w_{ij}) \right. \\ & \left. \quad - \frac{1}{2\sigma\tau_p^2} \frac{[x_{ij}-\mu_j-u_i^T v_j-\varphi_p w_{ij}]^2}{w_{ij}} - \frac{1}{\sigma} w_{ij} \right\}\,. \end{array}$$

The expectation is taken with respect to the hidden variables but conditionnally to the available sample and a current value of the parameters $\theta^{(t)}$. The resulting surrogate function for maximize iteratively instead of the former likelihood has the following analytical expression,

$$\begin{array}{rl} & \mathcal{Q}(\theta|\theta^{(t)}) \\ = & \mathbb{E}[\mathcal{L}(\theta, \mathbf{w}|x)|\theta^{(t)}, x] \\ = & \sum_{i,j} \left\{ -\frac{1}{2}\log(2\pi\tau_p^2\sigma^3) - \frac{1}{2}\mathbb{E}[\log(w_{ij})|\theta^{(t)}, x] \right. \\ & \left. \quad - \frac{1}{2\sigma\tau_p^2}\mathbb{E}[\frac{(x_{ij}-\mu_j-u_i^T v_j-\varphi_p w_{ij})^2}{w_{ij}}|\theta^{(t)}, x] - \frac{\mathbb{E}[w_{ij}|\theta^{(t)}, x]}{\sigma} \right. \\ = & \sum_{i,j} \left\{ C - \frac{(x_{ij}-\mu_j-u_i^T v_j-\varphi_p/e_{ij,-1})^2}{2\sigma\tau_p^2/e_{ij,-1}} - \frac{1}{\sigma} e_{ij,+1} \right. \\ & \left. \quad - \frac{\varphi_p^2(e_{ij,+1}-1/e_{ij,-1}^2)}{2\sigma\tau_p^2} - \frac{1}{2}\log(2\pi\tau_p^2\sigma^3) \right\}\,. \end{array}$$

The quadratic term is obtained by rewritting -see Appendix A- the obtained expression after developping all the terms in the square. The quantity $C$ is a constant not depending on the parameters.

- **Expectation-Maximization (EM):** The algorithm[29] leads to the maximization of the $\mathcal{Q}$ function as a new surrogate objective function because,

$$\log \mathcal{L}(\theta) \geq \mathcal{Q}(\theta|\theta^{(t)})\,.$$

Hence the maximization induces also the maximization of the former log-likelihood. This algorithm is compound of the two following steps.

*E-step:* The expectation step writes the expectation of the complete likelihood conditionally to the current value of the parameters and the sample. This involves several intermediate expectations with in particular the computation of the current values for $e_{ij,-1}$ and $e_{ij,+1}$.

*M-step:* The maximization step increases the resulting surrogate function $Q$ for same consequence on the likelihood. For light notation in the update equations, let denote the new intermediate cell data as follows,

$$y_{ij} = x_{ij} - \mu_j - \frac{\varphi_p}{e_{ij,-1}} \,.$$

The optimization for the parameters $\sigma$ and $\mu_j$, leads to solve for:

$$\frac{\partial Q(\theta|\theta^{(t)})}{\partial \sigma} = 0 \text{ and } \frac{\partial Q(\theta|\theta^{(t)})}{\partial \mu_j} = 0 \,.$$

The proof for the solutions are given in Appendix B while their expressions are in the main proposed algorithm in Figure 2 in Appendix C. The optimization for the parameters $\mathbf{u}$ and $\mathbf{v}$ is directly related to a factorization, the criterion can be rewritten with the help of the quantities $y_{ij}$. The corresponding optimisation problem ends to solve for all vectors $u_i$ and $v_j$ at the current step of the algorithm EM in order to get suitable update formula for:

$$(\mathbf{u},\mathbf{v}) \quad \leftarrow \quad \text{argmin}_{\mathbf{u},\mathbf{v}} F^{(t)}(\mathbf{u},\mathbf{v}) \,,$$

where, a matricial expression for the function above may be written with a product $\odot$,

$$\begin{aligned} F^{(t)}(\mathbf{u},\mathbf{v}) &= \textstyle\sum_{i,j} e_{ij,-1} \left( y_{ij}^{(t)} - u_i^T v_j \right)^2 \\ &= \| \mathbf{e}^{(t)} \odot (\mathbf{y}^{(t)} - \mathbf{u}^T \mathbf{v}) \|^2 \,. \end{aligned}$$

A closed-form solution is available from a generic algorithm for weighted factorizations or for regressions. Here, $\mathbf{e}^{(t)}$ stands for the matrix having for cells $e_{ij,-1}^{(t)}$ and $\mathbf{y}^{(t)}$ for the matrix with cell values equal to $y_{ij}^{(t)}$. Finally, the corresponding algorithm is summarized in Figure 2 in Appendix C, with a function for the weighted svd.

### Extension to pca lbm

The model and algorithm just above may be suitable for pca factorization of 2-d matrices in order to get an estimations of the parameters such that $\mathbf{u}$ or $\mathbf{v}$ need to be completed with an additional vector for a third tensor. For introducing a co-clustering and a tensor data, one may write that the classification log-likelihood can be written as previously, by changing the Gaussian distribution with a Laplace one:

$$L_C(\mathbf{z},\mathbf{w};\mathbf{x},\theta) \quad = \quad \log P(\mathbf{x}|\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3) + \sum_{\ell=1}^{3} \log P(\mathbf{z}_\ell) \,.$$

The different density and mass functions are:

$$\begin{aligned} P(\mathbf{z}) &= \prod_{\ell=1}^{3} \prod_{i_\ell,k_\ell} (p_{k_\ell})^{z_{i_\ell k_\ell}} \\ P(\mathbf{x}|\mathbf{z},\mathbf{w}) &= \prod_{I_m,\mathcal{K}_m} \varphi_L(x_{i_1 i_2 i_3};\mu_{k_1 k_2 k_3},\sigma_{k_1 k_2 k_3},p) \,. \end{aligned}$$

A solution for the parameters is found via variational expectation maximization after adding the constraints for the factorization, see [30] for a related variational em approach.

## 4 Discussion and perspectives

This document presents a multiway latent block model suitable for co-clustering tensor data, which is different to previous work [31] as more related to multivariate latent block models with different properties. A main interest herein is the reduction related to tensor factorizations and pca projections with any possible dimensions. Parameterization of latent block model is very powerfull, as it was able to lead to visualization of large datasets in [32] before and is expected to lead to a generic sparse model by selection of the blocks to be equal. This is written for two dimensional tensors and $r < k_1 \times k_2$ as follows:

$$\alpha_{k\ell} = \tau^T \alpha_{k\ell} \text{ with } \tau \in R^r \text{ and } \alpha_{k\ell} \in \{0,1\}^{\otimes r} \,.$$

Here $\tau$ is an unknown vector with continuous values, the binary vectors $\alpha_{k\ell}$ may be either constant [33] or either estimated for an automatic sparsity on the contrary to [34, 35] with non generic constraints. This is different to the latent block regression model [36] which was proposed after as another parameterization [13] for latent block models for an unrelated problem. Perspectives are alternative distributions and bayesian [37] or sequential estimations for big data, while alternative latent variables and additional constraints for better reductions are wanted.

## References

[1] A. K. Jain, "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, June 2010.

[2] J. A. Hartigan, *Clustering Algorithms*, Wiley, New York, 1975.

[3] G. Govaert, *Classification croisée*, Thèse d'état, Université Paris 6, France, 1983.

[4] G. Govaert, "Simultaneous clustering of rows and columns", *Control and Cybernetics*, vol. 24, no. 4, pp. 437–458, 1995.

[5] I. S. Dhillon, S. Mallela, and D. S. Modha, "Information-theoretic co-clustering", in *Proceedings of The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-2003)*, 2003, pp. 89–98.

[6] Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Srujana Merugu, and Dharmendra S. Modha, "A generalized maximum entropy approach to bregman co-clustering and matrix approximation", *J. Mach. Learn. Res.*, vol. 8, pp. 1919–1986, Dec. 2007.

[7] Deepak Agarwal and Srujana Merugu, "Predictive discrete latent factor models for large scale dyadic data", in *KDD*. 2007, pp. 26–35, ACM.

[8] Thomas Hofmann, "Probabilistic latent semantic analysis", *SIGIR'99*, pp. 50–57, 1999.

[9] Chris Ding, Tao Li, and Wei Peng, "On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing", *Computational Statistics and Data Analysis*, vol. 52, pp. 3913–3927, 2008.

[10] D. Cai, X. Wang, and X. He, “Probabilistic dyadic data analysis with local and global consistency”, in *ICML*, 2009.

[11] Gérard Govaert and Mohamed Nadif, “Clustering with block mixture models”, *Pattern Recognition*, vol. 36, no. 2, pp. 463–473, 2003.

[12] Gerard Govaert and Mohamed Nadif, “Latent block model for contingency table”, *Communications in Statistics-theory and Methods*, vol. 39, pp. 416–425, 2010.

[13] R. Priam and M. Nadif, “Data visualization via latent variables and mixture models: a brief survey”, *Pattern Analysis and Applications*, vol. 19, no. 3, pp. 807–819, Aug 2016.

[14] I.T. Jolliffe, *Principal Component Analysis*, Springer Verlag, 2002.

[15] A. P. Singh. and G. J. Gordon, “A unified view of matrix factorization models”, in *ECML PKDD, LNAI 5212*, 2008, pp. 358–373.

[16] L. Lebart, A. Morineau, and K. Warwick, *Multivariate Descriptive Statistical Analysis*, J. Wiley, 1984.

[17] Michael E. Tipping and Christopher M. Bishop, “Probabilistic principal component analysis”, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 3, no. 61, pp. 611–622, 1999.

[18] G.E. Hinton and R.R. Salakhutdinov, “Reducing the dimensionality of data with neural networks”, *Science (New York, N.Y.)*, vol. 313, pp. 504–7, 08 2006.

[19] J. Zhao and Q. Jiang, “Probabilistic pca for t distributions”, *Neurocomputing*, vol. 69, pp. 2217–2226, 2006.

[20] Cedric Archambeau, Nicolas Delannay, and Michel Verleysen, “Mixtures of robust probabilistic principal component analyzers”, *Neurocomputing*, vol. 71, no. 7-9, pp. 1274–1282, 2008.

[21] Junbin Gao, Paul W. Kwan, and Yi Guo, “Robust multivariate {L1} principal component analysis and dimensionality reduction”, *Neurocomputing*, vol. 72, no. 4-6, pp. 1242–1249, 2009.

[22] Mattias Villani and Rolf Larsson, “The multivariate split normal distribution and asymmetric principal components analysis”, *Communications in Statistics - Theory and Methods*, vol. 35, no. 6, pp. 1123–1140, 2006.

[23] Mia Hubert and Peter J. Rousseeuw, “Robpca: A new approach to robust principal component analysis”, *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.

[24] S. Engelen, M. Hubert, and K. Vanden Branden, “A comparison of three procedures for robust pca in high dimensions”, *Austrian Journal of Statistics*, vol. 34, pp. 117–126, 2005.

[25] Haipeng Shen Lingsong Zhang and Jianhua Z. Huang, “Robust regularized singular value decomposition with application to mortality data”, *Submitted to the Annals of Applied Statistics*, vol. 7, no. 3, pp. 1540–1561, 2013.

[26] Gerard Govaert and Mohamed Nadif, “An EM algorithm for the block mixture model”, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 643–647, 2005.

[27] Tamara G. Kolda and Brett W. Bader, “Tensor decompositions and applications”, *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.

[28] M. Nadif and G. Govaert, “Model-based co-clustering for continuous data”, in *ICMLA, IEEE Computer Society*, 2010, pp. 175–180.

[29] A.P. Dempster, N.M. Laird, and D.B. Rubin, “Maximum-likelihood from incomplete data via the EM algorithm”, *J. Royal Statist. Soc. Ser. B., 39*, pp. 1–38, 1977.

[30] Julie Aubert, Sophie Schbath, and Stéphane Robin, “Model-based biclustering for overdispersed count data with application in microbial ecology”, *Methods in Ecology and Evolution*, February 2021.

[31] Rafika Boutalbi, Lazhar Labiod, and Mohamed Nadif, “Tensor latent block model for co-clustering”, *International Journal of Data Science and Analytics*, vol. 10, no. 2, pp. 161–175, Aug 2020.

[32] R. Priam, M. Nadif, and G. Govaert, “Generalized topographic block model”, *Neurocomputing*, vol. 173, pp. 442–449, 2016.

[33] Rodolphe Priam, “Negative binomial latent block model with generalized constraints”, working paper or preprint, Mar. 2021.

[34] M. Ailem, F. Role, and M. Nadif, “Sparse poisson latent block model for document clustering”, *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1563–1576, 2017.

[35] Melissa Ailem, François Role, and Mohamed Nadif, “Model-based co-clustering for the effective handling of sparse data”, *Pattern Recognition*, vol. 72, pp. 108–122, 2017.

[36] Rafika Boutalbi, Lazhar Labiod, and Mohamed Nadif, “Latent block regression model”, 12 2023, pp. 73–81.

[37] Giulia Marchello, Marco Corneli, and Charles Bouveyron, “A deep dynamic latent block model for the co-clustering of zero-inflated data matrices”, in *Machine Learning and Knowledge Discovery in Databases: Research Track*, Cham, 2023, pp. 695–710, Springer Nature Switzerland.

# Appendix A

Let's have $w$ a random variable and $a$, $b$ two deterministic constants. Let's also have $e_{+1} = E(w)$ and $e_{-1} = E(w^{-1})$. Then, it can be written,

$$\begin{aligned} & \mathbb{E}[w^{-1}(a-bw)^2] \\ = \; & \mathbb{E}[w^{-1}]a^2 - 2ab + b^2\mathbb{E}[w] \\ = \; & e_{-1}[a - b/e_{-1}]^2 + b^2 e_{+1} - (b/e_{-1})^2 \\ = \; & \tfrac{(a-b/e_{-1})^2}{1/e_{-1}} + b^2(e_{+1} - 1/e_{-1}^2)\,. \end{aligned}$$

# Appendix B

$$\begin{aligned} & \tfrac{\partial Q(\theta|\theta^{(t)})}{\partial \mu_j} = 0 \\ & \to \tfrac{\partial}{\partial \mu_j} \textstyle\sum_i \Big\{ -\tfrac{(x_{ij}-\mu_j-u_i^T v_j-\varphi_p/e_{ij,-1})^2}{2\sigma\tau_p^2/e_{ij,-1}} - \tfrac{1}{\sigma} e_{ij,+1} \\ & \qquad -\tfrac{\varphi_p^2(e_{ij,+1}-1/e_{ij,-1}^2)}{2\sigma\tau_p^2} - \tfrac{1}{2}\log(2\pi\tau_p^2\sigma^3) \Big\} = 0 \\ & \to \tfrac{\partial}{\partial \mu_j} \textstyle\sum_i e_{ij,-1} \left\{ -(x_{ij} - \mu_j - u_i^T v_j - \varphi_p/e_{ij,-1})^2 \right\} = 0 \\ & \to -\textstyle\sum_i e_{ij,-1}\mu_j + \sum_i e_{ij,-1} \left\{ (x_{ij} - u_i^T v_j - \varphi_p/e_{ij,-1})^2 \right\} = 0 \\ & \to \mu_j = \tfrac{\sum_i e_{ij,-1} \sum_i e_{ij,-1} \left\{ (x_{ij}-u_i^T v_j-\varphi_p/e_{ij,-1})^2 \right\}}{\sum_i e_{ij,-1}}{}_{,}\,. \end{aligned}$$

$$\begin{aligned} & \tfrac{\partial Q(\theta|\theta^{(t)})}{\partial \sigma} = 0 \\ & \to \tfrac{\partial}{\partial \sigma} \textstyle\sum_{i,j} \Big\{ -\tfrac{(x_{ij}-\mu_j-u_i^T v_j-\varphi_p/e_{ij,-1})^2}{2\sigma\tau_p^2/e_{ij,-1}} - \tfrac{1}{\sigma} e_{ij,+1} \\ & \qquad -\tfrac{\varphi_p^2(e_{ij,+1}-1/e_{ij,-1}^2)}{2\sigma\tau_p^2} - \tfrac{1}{2}\log(2\pi\tau_p^2\sigma^3) \Big\} = 0 \\ & \to \textstyle\sum_{i,j} \Big\{ +\tfrac{(x_{ij}-\mu_j-u_i^T v_j-\varphi_p/e_{ij,-1})^2}{2\sigma^2\tau_p^2/e_{ij,-1}} + \tfrac{1}{\sigma^2} e_{ij,+1} \\ & \qquad +\tfrac{\varphi_p^2(e_{ij,+1}+1/e_{ij,-1}^2)}{2\sigma^2\tau_p^2} - \tfrac{3}{2}\tfrac{1}{\sigma} \Big\} = 0 \\ & \to \textstyle\sum_{i,j} \tfrac{2\sigma^2}{3} \Big\{ +\tfrac{e_{ij,-1}(y_{ij}-u_i^T v_j)^2+\varphi_p^2(e_{ij,+1}+1/e_{ij,-1}^2)}{2\sigma^2\tau_p^2} + \tfrac{1}{\sigma^2} e_{ij,+1} \Big\} \\ & \qquad \tfrac{2\sigma^2}{3} \textstyle\sum_{i,j} - \tfrac{3}{2}\tfrac{1}{\sigma} = 0 \\ & \to \textstyle\sum_{i,j} \Big\{ +\tfrac{e_{ij,-1}(y_{ij}-u_i^T v_j)^2+\varphi_p^2(e_{ij,+1}+1/e_{ij,-1}^2)}{3\tau_p^2} + \tfrac{2}{3\sigma^2} e_{ij,+1} \Big\} \\ & \qquad -\sigma \times [\textstyle\sum_{i,j} 1] = 0 \\ & \to \textstyle\sum_{i,j} \Big\{ +\tfrac{e_{ij,-1}(y_{ij}-u_i^T v_j)^2+\varphi_p^2(e_{ij,+1}+1/e_{ij,-1}^2)}{3\tau_p^2} + \tfrac{2}{3\sigma^2} e_{ij,+1} \Big\} \\ & \qquad = \sigma \times m \times n = 0 \end{aligned}$$

# Appendix C

```
0) Initialization:
Initialize $\{\mu_j^{(0)}\}$, $\sigma^{(0)}$, $\mathbf{u}^{(0)}$ and $\mathbf{v}^{(0)}$.

1) E-Step:
Update $\{e_{ij,+1}^{(t)}\}$ by (3), and $\{e_{ij,-1}^{(t)}\}$ by (3).

2) M-Step:
Update for $1 \leq j \leq m$ every component of $\mu$, and $\sigma$, and:

    $\sigma^2 \leftarrow \frac{1}{nm}\sum_{i,j}\left\{ \frac{\left[e_{ij,-1}\left(y_{ij}-u_i^T v_j\right)^2+\varphi_p^2(e_{ij,+1}-1/e_{ij,-1}^2)\right]}{3\tau_p^2} + \frac{2e_{ij,+1}}{3} \right\}$
    $\mu_j \leftarrow \frac{1}{\sum_i e_{ij,-1}} \sum_i e_{ij,-1} \left( x_{ij} - u_i^T v_j - \frac{\varphi_p}{e_{ij,-1}} \right)$

  Update $\mathbf{u}$ and $\mathbf{v}$, with:

    $(\mathbf{u},\mathbf{v}) \leftarrow wsvd(\mathbf{y}^{(t)},\mathbf{e}^{(t)})$.

3) End:
If $\frac{\left|Q(\theta^{(t+1)}|\theta^{(t)})-Q(\theta^{(t)}|\theta^{(t-1)})\right|}{\left|Q(\theta^{(t)}|\theta^{(t-1)})\right|} < \varepsilon$ then stop else return to 1).
```

Figure 1: Algorithm for learning the parameters of qPCA using a function $wsvd = wsvd(A,W)$ for a weighted singular decomposition of the matrix A as a first entry with weigths W as a second entry.

```
0) Initialization:
Initialize $\mathbf{u}^{(0)}$ and $\mathbf{v}^{(0)}$.

0) Update u and v:

  - Update the rows of $\mathbf{u}$, for $1 \leq i \leq n$ .

      $u_i^{t+1} \leftarrow \mathrm{argmin}_{u_i} \parallel x_i - \mathbf{v}^{(t)} u \parallel_{\mathbf{w}_i}^2$ .

  - Update the rows of $\mathbf{v}$, for $1 \leq j \leq p$ .

      $v_j^{t+1} \leftarrow \mathrm{argmin}_{v_j} \parallel x^j - \mathbf{u}^{(t)} v \parallel_{\mathbf{w}^j}^2$ .

3) End:
If $\frac{\|\mathbf{u}^{(t+1)}-\mathbf{u}^{(t)}\|}{\|\mathbf{u}^{(t)}\|} + \frac{\|\mathbf{v}^{(t+1)}-\mathbf{v}^{(t)}\|}{\|\mathbf{v}^{(t)}\|} < \varepsilon$ then stop else return to 1).
```

Figure 2: Algorithm for the function $wsvd = wsvd(\mathbf{x},\mathbf{w})$ for a weighted singular decomposition of the matrix A as a first entry with weigths W as a second entry.