Model-based Co-clustering for High Dimensional Sparse Data

Supplementary Material

Aghiles Salah
aghiles.salah@parisdescartes.fr

Nicoleta Rogovschi
nicoleta.rogovschi@parisdescartes.fr

Mohamed Nadif
mohamed.nadif@parisdescartes.fr

LIPADE, University of Paris Descartes
45 Rue des Saints-Peres,
75006, Paris, France

Appendix A. Parameters Estimation

In this appendix, we provide the derivation details of the maximum likelihood estimates for parameters of the proposed model dbmovMFs.

A.1 Maximum Likelihood Estimate

The expectation of the complete data log-likelihood is given by

$$\begin{aligned} E[L_c(\Theta|\mathcal{X},\mathbf{Z})] &= \sum_h \tilde{z}_{.h}\log\alpha_h + \sum_h \tilde{z}_{.h}\log(c_d(\kappa_h)) \\ &+ \sum_{h,i,j} \tilde{z}_{ih} w_{jh}\kappa_h\mu_{hh}x_{ij} \qquad (1) \end{aligned}$$

where $\tilde{z}_{.h} = \sum_i \tilde{z}_{ih}$. We first maximize (1) with respect to α_h, subject to the constraint $\sum_h \alpha_h = 1$. The corresponding Lagrangian, up to terms which are not function of α_h, is given by

$$L(\boldsymbol{\alpha},\lambda) = \sum_h \tilde{z}_{.h}\log\alpha_h + \lambda_h(1 - \sum_h \alpha_h) \qquad (2)$$

Taking derivatives with respect to α_h's, we obtain

$$\frac{\partial L(\boldsymbol{\alpha},\lambda)}{\partial\alpha_h} = \frac{\tilde{z}_{.h}}{\alpha_h} - \lambda$$

setting this derivative to zero yields:

$$\tilde{z}_{.h} = \lambda\alpha_h$$

Summing both sides over all h yields to $\lambda = n$, thereby the maximizing value of the parameter α_h is given by:

$$\hat{\alpha}_h = \frac{\tilde{z}_{.h}}{n} \qquad (3)$$

In the same manner, to maximize expectation (1) with respect to $\boldsymbol{\mu}_h^{\mathbf{w}}$, subject to the constraint $(\boldsymbol{\mu}_h^{\mathbf{w}})^T\boldsymbol{\mu}_h^{\mathbf{w}} = 1$, we form the corresponding Lagrangian by isolating the terms which depends on $\boldsymbol{\mu}_h^{\mathbf{w}}$, which leads to

$$L(\boldsymbol{\mu},\lambda) = \sum_{h,i,j} \tilde{z}_{ih} w_{jh}\kappa_h\mu_{hh}x_{ij} + \lambda_h(1 - \sum_j w_{jh}\mu_{hh}^2)$$

Taking the derivative with respect to μ_{hh}, yields:

$$\frac{\partial L(\boldsymbol{\mu},\lambda)}{\partial\mu_h} = \sum_{i,j} \tilde{z}_{ih} w_{jh}\kappa_h x_{ij} - 2\lambda w_{.h}\mu_{hh}$$

where $w_{.h} = \sum_j w_{jh}$. Setting this derivative to zero, we obtain:

$$\lambda\mu_{hh} = \frac{\sum_{i,j}\tilde{z}_{ih} w_{jh}\kappa_h x_{ij}}{2w_{.h}}$$

Thus,

$$\lambda^2\mu_{hh}^2 = \frac{(\sum_{i,j}\tilde{z}_{ih} w_{jh}\kappa_h x_{ij})^2}{4w_{.h}^2}$$

Multiplying both sides by $w_{.h}$, yields:

$$\lambda^2 w_{.h}\mu_{hh}^2 = \frac{(\sum_{i,j}\tilde{z}_{ih} w_{jh}\kappa_h x_{ij})^2}{4w_{.h}} \qquad (4)$$

hence, we obtain

$$\begin{aligned} \lambda &= \kappa_h\frac{\sqrt{w_{.h}(\sum_{i,j}\tilde{z}_{ih} w_{jh} x_{ij})^2}}{2w_{.h}} \\ &= \kappa_h\frac{\|\mathbf{r}_h^{\mathbf{w}}\|}{2w_{.h}} \end{aligned}$$

where $\mathbf{r}_h^{\mathbf{w}}$ is a d dimensional vector: let $j' = 1,\ldots,d$, $r_{hj'}^{\mathbf{w}} = r_h^{\mathbf{w}} = \sum_{i,j}\tilde{z}_{ih} w_{jh} x_{ij}$ if $w_{jh} = 1$ and $r_{hj'}^{\mathbf{w}} = 0$, otherwize. Hence, the maximizing value of the parameter μ_{hh} is given by:

$$\begin{aligned}\hat{\mu}_{hh} &= \frac{\sum_{i,j}\tilde{z}_{ih}w_{jh}x_{ij}}{\|\mathbf{r}_h^{\mathbf{w}}\|} \\ &= \frac{\sum_{i,j}\tilde{z}_{ih}w_{jh}x_{ij}}{\sqrt{w_{.h}(\sum_{i,j}\tilde{z}_{ih}w_{jh}x_{ij})^2}} \\ &= \pm\frac{1}{\sqrt{w_{.h}}} \qquad (5)\end{aligned}$$

according to whether $r_h^{\mathbf{w}} = \sum_{i,j}\tilde{z}_{ih}w_{jh}x_{ij}$ is positive or negative. It follows from equation (5) that given the column partition $\mathbf{w}$ and the sign of $r_h^{\mathbf{w}}$, the centroid parameter $\boldsymbol{\mu}_h^{\mathbf{w}}$ can be deduced directly.

Next we concentrate on maximizing equation (1), with respect to the concentration parameters κ_h's, subject to the constraint $\kappa_h > 0$, $\forall h$. The Lagrangian up to terms which do not contains κ_h is given by

$$L(\kappa) = \sum_h \tilde{z}_{.h}\log(c_d(\kappa_h)) + \sum_{h,i,j}\tilde{z}_{ih}w_{jh}\kappa_h\hat{\mu}_{hh}x_{ij} \qquad (6)$$

note that, by KKT conditions, the Lagrangian multiplier for the constraint $\kappa_h > 0$ has to be equal to zero. Taking the partial derivative of equation (6) with respect to κ_h, we obtain

$$\frac{\partial L(\kappa)}{\partial\kappa_h} = \tilde{z}_{.h}\frac{c'_d(\kappa_h)}{c_d(\kappa_h)} + \sum_{i,j}\tilde{z}_{ih}w_{jh}\hat{\mu}_{hh}x_{ij}$$

Setting this derivative equal to zero, leads to:

$$\frac{c'_d(\kappa_h)}{c_d(\kappa_h)} = -\frac{\hat{\mu}_{hh}\times\sum_{i,j}\tilde{z}_{ih}w_{jh}x_{ij}}{\tilde{z}_{.h}}$$

replacing $\hat{\mu}_{hh}$ by $\frac{\sum_{i,j}\tilde{z}_{ih}w_{jh}x_{ij}}{\|\mathbf{r}_h^{\mathbf{w}}\|}$ (see, equation 5), we obtain:

$$\frac{c'_d(\kappa_h)}{c_d(\kappa_h)} = -\frac{\|\mathbf{r}_h^{\mathbf{w}}\|}{\tilde{z}_{.h}\hat{w}_{.h}}$$

let $s = d/2 - 1$, then:

$$\begin{aligned}c'_d(\kappa_h) &= \frac{s\kappa_h^{s-1}(2\pi)^{s+1}I_s(\kappa_h) - \kappa_h^s(2\pi)^{s+1}I'_s(\kappa_h)}{(2\pi)^{2s+2}I_s^2(\kappa_h)} \\ &= \frac{s\kappa_h^{s-1}}{(2\pi)^{s+1}I_s(\kappa_h)} - \frac{\kappa_h^s I'_s(\kappa_h)}{(2\pi)^{s+1}I_s^2(\kappa_h)} \\ &= c_d(\kappa_h)\left(\frac{s}{\kappa_h} - \frac{I'_s(\kappa_h)}{I_s(\kappa_h)}\right) \qquad (7)\end{aligned}$$

Hence,

$$\frac{-c'_d(\kappa_h)}{c_d(\kappa_h)} = \frac{I_{s+1}(\kappa_h)}{I_s(\kappa_h)} = \frac{I_{d/2}(\kappa_h)}{I_{d/2-1}(\kappa_h)} \qquad (8)$$

The above equation (8), arises from the use of the following recurrence formula [Abramowitz and Stegun, 1964]:

$$\kappa_h I_{s+1}(\kappa_h) = \kappa_h I'_s(\kappa_h) - sI_s(\kappa_h) \qquad (9)$$

Note that computing the maximizing value $\hat{\kappa}_h$ from equation (7) implies to inverse a ratio of Bessel function, a problem for which there is no closed-form solution. Thus, Following Banerjee et al. [2005], we can derive an accurate approximation of the concentration parameter, by using the following continued fraction formula:

$$\frac{I_{d/2}(\kappa_h)}{I_{d/2-1}(\kappa_h)} = \frac{1}{\frac{d}{\kappa_h} + \frac{1}{\frac{d+2}{\kappa_h}+\ldots}}. \qquad (10)$$

Letting $\bar{r}_h^{\mathbf{w}} = \frac{\|\mathbf{r}_h^{\mathbf{w}}\|}{\tilde{z}_{.h}\hat{w}_{.h}} = \frac{I_{d/2}(\kappa_h)}{I_{d/2-1}(\kappa_h)}$ and using equation (10), we obtain:

$$\frac{1}{\bar{r}_h^{\mathbf{w}}} \approx \frac{d}{\kappa_h} + \bar{r}_h^{\mathbf{w}}$$

which yields the following approximation:

$$\hat{\kappa}_h = \frac{d\bar{r}_h^{\mathbf{w}}}{1 - (\bar{r}_h^{\mathbf{w}})^2}$$

Finally, the authors in [Banerjee et al., 2005] have empirically shown that adding the following correction term $\frac{-(\bar{r}_h^{\mathbf{w}})^3}{1-(\bar{r}_h^{\mathbf{w}})^2}$ results in a better approximation of $\hat{\kappa}_h$, which leads to:

$$\hat{\kappa}_h = \frac{d\bar{r}_h^{\mathbf{w}} - (\bar{r}_h^{\mathbf{w}})^3}{1 - (\bar{r}_h^{\mathbf{w}})^2} \qquad (11)$$

As opposed to the classical movMFs where it is easy to verify that $\bar{r}_h \leq 1$ (see equation 6*c*) given the definition of $\mathbf{r}$, it is not straightforward to verify that $\bar{r}_h^{\mathbf{w}} \leq 1$, without careful analysis. Such a result is imperative, to guarantee that the concentration parameters are positive, i.e, $\kappa_h > 0$, $\forall h$, specially when using the approximation of equation (11). Hence, the following Proposition provides theoretical guarantee about the fact that $0 \leq \bar{r}_h^{\mathbf{w}} \leq 1$, thereby it ensures that κ_h estimated from equation (11) is always positive.

Proposition 1 *Let* $\mathbf{r}$ *be a non-zero vector in* $\mathbb{R}^d$ *(i.e.,* $\mathbf{r} = (r_1, \ldots, r_d)^T$*, such as* $d \geq 2$*) which results from a weighted sum of n d-dimensional unit vector, i.e,* $\mathbf{r} = \sum_i p_i\mathbf{x}_i$*,* $\mathbf{x}_i \in \mathbb{R}^d$ *and* $\|\mathbf{x}_i\| = 1$ *,* $\forall i \in \{1, \ldots, n\}$ *,* $n \geq 2$*, the weights* $p_i \geq 0$*,* $\forall i$*. Let* $\mathbf{r}^d$ *be a vector in* $\mathbb{R}^d$*, such* *as all its components are equal to the sum of elements of* $\mathbf{r}$ *(i.e,* $\mathbf{r}^d_1 = \cdots = \mathbf{r}^d_d = \sum_{j=1}^d r_j$*). Then* $0 < \|\mathbf{r}^d\| \leq d \times \sum_i p_i$ *with equality only if all unit vectors* $\mathbf{x}_i$ *are equal/collinear.*

Proof. We define two vectors $\mathbf{d}$ and $\mathbf{r}^+$ in $\mathbb{R}^d$ as follows: $\mathbf{d} = \frac{1}{\sqrt{d}}\mathbb{1}$ and $r^+_j = |r_j|, \quad \forall j \in \{1, \ldots, d\}$. as $\|\mathbf{r}^d\|$, d and $\|\mathbf{r}\|$ are all positive, we aim to show that $\frac{\|\mathbf{r}^d\|}{d \times \sum_i p_i} \leq 1$, we have:

$$\begin{aligned}
\frac{\|\mathbf{r}^d\|}{d} &= \frac{\sqrt{(r^d_1)^2 + \ldots + (r^d_d)^2}}{d} \\
&= \frac{\sqrt{d \times \left(\sum_j r_j\right)^2}}{d} \\
&= \frac{\sqrt{d} \times \left|\sum_j r_j\right|}{d} \\
&= \frac{1}{\sqrt{d}} \times \left|\sum_j r_j\right| \\
&\leq \frac{1}{\sqrt{d}} \times \sum_j |r_j| \\
&\leq \mathbf{d}^t.\mathbf{r}^+ \\
&\leq \|\mathbf{d}\|\|\mathbf{r}^+\| \cos(\mathbf{d}, \mathbf{r}^+)
\end{aligned}$$

by definition of $\mathbf{r}^+$ and $\mathbf{d}$, we have $\|\mathbf{r}^+\| = \|\mathbf{r}\|$ thereby $\|\mathbf{r}^+\| \leq \sum_i p_i$ (i.e, $\|\mathbf{r}^+\| = \|\mathbf{r}\| = \|p_1\mathbf{x}_1 + \cdots + p_n\mathbf{x}_n\| \leq \|p_i\mathbf{x}_1\| + \cdots + \|p_n\mathbf{x}_n\| = \sum_i p_i$) and $\|\mathbf{d}\| = 1$, hence

$$\frac{\|\mathbf{r}^d\|}{d} \leq \sum_i p_i \times \cos(\mathbf{d}, \mathbf{r}^+) \tag{12}$$

dividing both sides by $\sum_i p_i$, we get

$$\frac{\|\mathbf{r}^d\|}{d \times \sum_i p_i} \leq \cos(\mathbf{d}, \mathbf{r}^+) \tag{13}$$

by definition both $\mathbf{d}$ and $\mathbf{r}^+$ are non-zero vectors and lie on the first orthant of d-dimensional unit hypersphere, thus,

$$0 < \frac{\|\mathbf{r}^d\|}{d \times \sum_i p_i} \leq \cos(\mathbf{d}, \mathbf{r}^+) \leq 1$$

The equality holds only if $\mathbf{d}$ and $\mathbf{r}^+$ are collinear, thereby all components of $\mathbf{r}$ are equal.

Appendix B. Experiments

B.1 Evaluation measures

In this appendix we give some details about the clustering-evaluation measures—Normalized Mutual Information NMI and Ajusted Rand Index ARI—used in our experiments. The NMI is estimated as follows

$$NMI = \frac{\sum_{k\ell} \pi_{h\ell} \log \frac{\pi_{h\ell}}{\pi_h \hat{\pi}_\ell}}{\sqrt{(\sum_h \pi_h \log \pi_h)(\sum_\ell \hat{\pi}_\ell \log \hat{\pi}_\ell)}}$$

where π_h denotes the proportion of elements in the resulting cluster h, while $\hat{\pi}_l$ denotes the proportion of class (true cluster) ℓ, i.e, $\pi_h = n_h/n$, $\hat{\pi}_\ell = \hat{n}_\ell/n$; n, n_h and $\hat{n}_\ell$ denote the total number of objects, the number of objects in cluster h and the number of objects in class ℓ, respectively. The proportion of objects that are common to cluster h and class ℓ is denoted by $\pi_{h\ell}$. Intuitively NMI quantifies how much the estimated clustering is informative about the true clustering, it can be shown that NMI lies in the range $[0, 1]$. If the resulting clustering and the true one are identical, then NMI = 1. However, when the obtained clusters are substantially different from the true classes then the value of the NMI will be low and close to zero for a random clustering.

The ARI measures the correspondence between two clusterings. As it has been demonstrated by Milligan and Cooper [1986], ARI is a superior measure compared to several other measures, for assessing the correspondence between two clusterings. Formally, the ARI is given by

$$ARI = \frac{\sum_{h,\ell} \binom{n_{h\ell}}{2} - \sum_h \binom{n_{h.}}{2} \sum_\ell \binom{n_{.\ell}}{2} / \binom{n}{2}}{\frac{1}{2}\left[\sum_h \binom{n_{h.}}{2} + \sum_\ell \binom{n_{.\ell}}{2}\right] - \sum_h \binom{n_{h.}}{2} \sum_\ell \binom{n_{.\ell}}{2} / \binom{n}{2}}$$

where $n_{h.}$, $n_{.\ell}$, $n_{h\ell}$ denote respectively the number of objects in cluster h, in class ℓ, that are in cluster h as well as in class ℓ.

Intuitively, the ARI measures the degree of agreement between an estimated clustering and a reference clustering. Hence, ARI = 1 if the estimated clustering and the true one agree perfectly, and ARI is close to zero for random clustering.

References

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables.* Number 55. Courier Corporation, 1964.

Arindam Banerjee, Inderjit S. Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. *J. Mach. Learn. Res.*, 6:1345–1382, 2005.

Glenn W Milligan and Martha C Cooper. A study of the comparability of external criteria for hierarchical cluster analysis. *Multivariate Behavioral Research*, 21(4):441–458, 1986.