Simultaneous Dimension Reduction and Clustering via the NMF-EM Algorithm

Léna Carel
Transdev and CREST, ENSAE, Université Paris Saclay
and
Pierre Alquier
CREST, ENSAE, Université Paris Saclay

June 7, 2018

Abstract

Mixture models are among the most popular tools for clustering. However, when the dimension and the number of clusters is large, the estimation of the clusters become challenging, as well as their interpretation. Restriction on the parameters can be used to reduce the dimension. An example is given by mixture of factor analyzers for Gaussian mixtures. The extension of MFA to non-Gaussian mixtures is not straightforward. We propose a new constraint for parameters in non-Gaussian mixture model: the K components parameters are combinations of elements from a small dictionary, say H elements, with $H \ll K$. Including a nonnegative matrix factorization (NMF) in the EM algorithm allows us to simultaneously estimate the dictionary and the parameters of the mixture. We propose the acronym NMF-EM for this algorithm, implemented in the R package `nmfem`. This original approach is motivated by passengers clustering from ticketing data: we apply NMF-EM to data from two Transdev public transport networks. In this case, the words are easily interpreted as typical slots in a timetable.

$<$ *Keywords:* Mixture models, ticketing data, matrix factorization, reduction of dimension, EM algorithm, clustering, hidden variables.

1 Introduction

With the growing ability to collect and store data in transports system, electricity consumption and more, urban computing is becoming a major tool in urban policy and planning [57]. For example, for transports system, there is a growing litterature on ticketing and smart-card data processing in trains and buses [39, 42, 16, 44, 11, 51], bike-sharing systems [46, 14, 9, 23] or taxis [43]. Our objective in this paper is to propose a clustering method for users, and for stations, that would be adapted to ticketing data collected by Transdev, a public network company. This method could be suitable for clustering structured high-dimensional data in other applications.

The range of machine learning and statistical tools used in urban computing is large. This goes from descriptive data-mining techniques as in [39] to statistical models as in [16]. The model-based clustering approach in [16] is actually close to our objective: journeys of a user are seen as realizations of multinomials random variables. The parameters of these distributions depends of the user only through the cluster the user belongs to. The complete model for journeys is thus a mixture of multinomials. The authors estimate the parameters and the clusters by the EM algorithm (see Chapter 9 in [6] for an introduction; many R packages are available, `mcclust` [47] is extremely complete for clustering with Gaussian mixtures, `mixtools` [4] is a more generalist package covering other distributions, including multinomials). Model-based clustering was also used for transport data in [14, 9] with nice results. However, there are some issues with this approach. When the dimension is large, the estimates are likely to have a large variance (curse of dimensionality). It might also be difficult to interpret clusters described by a huge number of parameters: it is indeed argued in [11] that some profiles in [16] are not easily interpretable. It seems then necessary to reduce the dimension, that is, to impose some restrictions on the parameters that will reduce the variance and increase the interpretability.

Since the seminal work on model-based clustering [52], various examples of such restrictions have been proposed. We refer the reader to [19, 8, 35, 34, 22] for recent surveys on existing approaches (see also [32, 12] for a more general overview on mixtures). A first approach is variable selection [45]. This method is now well understood from an empirical perspective [49] as well as from a theoretical point of view [30, 31]. See [13] for more recent advances and [18] for a nice survey. The underlying assumption is that clusters differ only through a few variables. This assumption is satisfied in many examples presented in the aforementionned papers. However, it does not seem to be adapted to our case. The difference between two users, say a student and a retired person, is that the student has a regular travel schedule, while the retired person usually doesn't. This is a typical example of a strong structure that is not summarized by a small number of variables. Another approach for dimension reduction in mixtures is the mixture of factor analyzer (MFA) introduced in [20, 33], see [38, 36, 40] for recent extensions. In MFA, the means and variances depends on the cluster, and the variance might be concentrated in some directions. This is more related to our objective, but this model was developed for mixture of Gaussians. Travels patterns are modeled by mixture of multinomials in [16].

In this paper, we propose a new model that can be seen as an adaptation of MFA to mixture

of distributions with nonnegative parameters (including multinomial distributions). The decomposition in Gaussian factors in MFA is replaced by a nonnegative matrix factorization (NMF). Introduced by [26], NMF rewrites columns of a given matrix with nonnegative entries as combinations of elements in a small dictionary. These elements are often refered to as "words". These words play a somewhat similar role to factors in MFA, even though the formalism is different. For example these words are not modelled as random variables. We provide an adaptation of the celebrated EM algorithm to this setting. We refer to this algorithm as NMF-EM. It is available as an R package, `nmfem`.

The paper is organized as follows. In Section 2 we describe our model and the general form of NMF-EM. Motivated by the ticketing data, we provide the detailed form of the algorithm in the case of mixture of multinomials (Subsection 2.3). The clustering abilities of NMF-EM are compared to the ones of EM (without reduction of dimension) and of k-means in a short simulation study in Section 3. We finally present results on ticketing data provided by the Transdev Group in Section 4 (more details on this real data study can be found in the supplementary material).

2 Factorization of mixture parameters and the NMF-EM algorithm

2.1 Factorization of mixture parameters

Given a parametric family of distributions $(f_\vartheta)_{\vartheta \in \mathbb{R}^M}$, assume the observations $Y_1, \ldots, Y_n$ are i.i.d from

$$\sum_{k=1}^K p_k f_{\theta_{\cdot,k}}(\cdot), \tag{1}$$

where each $\theta_{\cdot,k} \in \mathbb{R}^M$ is a column of a $K \times M$ matrix θ. For the sake of brevity, let $p = (p_1, \ldots, p_K)$, which belongs to the simplex $\mathcal{S}_K = \{\rho \in \mathbb{R}_+^K : \rho_1 + \cdots + \rho_K = 1\}$. A way to rephrase this mixture model which is useful for clustering purposes is to introduce i.i.d hidden class variables: $Z_i = (Z_{i,1}, \ldots, Z_{i,K}) \sim \mathcal{M}ult(p, 1)$. Here, $\mathcal{M}ult(p, 1)$ denotes the multinomial distribution, that is, the probability that Z_i is the k-th basis vector $(0, \ldots, 1, \ldots, 0)$ is given by p_k. Taking $Y_i \big| (Z_{i,k} = 1) \sim f_{\theta_{\cdot,k}}(\cdot)$ implies that the Y_i's are actually i.i.d from (1).

In model-based clustering, estimation of the Z_i's allow to assign each Y_i to a cluster k while the estimation of $\theta_{\cdot,k}$ provides a summary of the information on location, scale and shape of cluster k. Still, as argued in the introduction, when the dimension M is too large, this information can be unreliable and uneasy to interpret. Many dimension reduction methods were proposed, among them MFA for mixtures of Gaussians. A standard mixture of Gaussian in $\mathbb{R}^d$ is $Y_i \big| (Z_{i,k} = 1) \sim \mathcal{N}(\mu_k, \Sigma_k)$, the simplest form of MFA is given by $Y_i \big| (Z_{i,k} = 1) \sim \mathcal{N}(\mu_k, \Lambda_k \Lambda_k^T + \sigma^2 I)$, where Λ_k is a $d \times H$ matrix with $H \ll d$. Thus, the estimation of the $d \times d$ matrix Σ_k is reduced to the estimation of the much smaller $H \times d$ matrix Λ_k. An interpretation of this model is that Y_i depends not only on the hidden variable Z_i but also on hidden factors $X_i \sim \mathcal{N}(0, I_H)$:

$\mathbb{E}(Y_i | X_i = x, Z_{i,k} = 1) = \Lambda_k x + \mu_k$. See the references given in the introduction, e.g Section 5 in [8]. This model provides reduction of dimension and has a nice interpretation, but is is not direct to extend it beyond Gaussian variables.

In the case where of multinomial distributions, and more generally in the case where the parameters ϑ of $(f_\vartheta)_{\vartheta \in \mathbb{R}^M}$ are actually nonnegative, one could think of restrictions on the mixture parameters matrix θ that would similarly involve a small number H of hidden factors. But one has to be careful: Gaussian hidden factors would in general not generate nonnegative parameters. In a celebrated paper [26], Lee and Seung proposed a dimension reduction tool for matrices with nonnegative entries: NMF (nonnegative matrix factorization). The idea is to factorize a $K \times M$ matrix θ as

$$
\underbrace{\begin{pmatrix} \theta_{1,1} & \dots & \theta_{1,K} \\ \vdots & \ddots & \vdots \\ \theta_{M,1} & \dots & \theta_{M,K} \end{pmatrix}}_{\theta} = \underbrace{\begin{pmatrix} \Phi_{1,1} & \dots & \Phi_{1,H} \\ \vdots & \ddots & \vdots \\ \Phi_{M,1} & \dots & \Phi_{M,H} \end{pmatrix}}_{\Phi} \underbrace{\begin{pmatrix} \Lambda_{1,1} & \dots & \Lambda_{1,K} \\ \vdots & \ddots & \vdots \\ \Lambda_{H,1} & \dots & \Lambda_{H,K} \end{pmatrix}}_{\Lambda} \tag{2}
$$

with $H \leq K, M$, under the assumption that all the entries in Φ and Λ are nonnegative. When $H \ll K, M$, the dimension reduction is substantial. NMF rewrites columns of a given matrix as positive combinations of elements, or words, in a small dictionary Λ. It turns out that this dictionary is often easily interpretable. NMF was succesfully used as a data mining tool in document clustering [54, 48], collaborative filtering and recommender systems on the Web [25, 29], dictionary learning for images [26], topic extraction in texts [41] or time series recovering [37], among others. It was also used as a data mining tool for transports data by [23, 43, 44, 51] and our previous work [11]: we "compressed" the data $Y_1, \dots, Y_n$ using an NMF and then used a (model-free) clustering algorithm on the compressed observations. The improvement in terms of interpretability with respect to [16] was substantial. However, this approach was completely *ad hoc*: there are many possible criterion to approximate NMF: the Poisson-likekihood [26, 27], the quadratic criterion or Gaussian-likelihood [10, 27], the Ikuro-Saito divergence [17]... In a model-free approach, the choice of the criterion is difficult. The mixture model (1) leads to a natural criterion: the likelihood.

We are finally in position to define our model: we use NMF as a restriction on nonnegative parameters in mixture models. That is, $Y_1, \dots, Y_n$ are i.i.d from

$$
g_{p,\Phi,\Lambda}(\cdot) = \sum_{k=1}^K p_k f_{(\Phi\Lambda)._{,k}}(\cdot) \tag{3}
$$

or equivalently, $Y_i | (Z_{i,k} = 1)$ is drawn from $f_{(\Phi\Lambda)._{,k}}$ and $Z_i \sim \mathcal{Mult}(p, 1)$. The model is parametrized by $p \in \mathcal{S}_K$, $\Lambda \in \mathbb{R}_+^{M \times H}$ and $\Phi \in \mathbb{R}_+^{H \times K}$. For short, put $Y = (Y_1, \dots, Y_n)$ and $Z = (Z_1, \dots, Z_n)$. The log-likelihood is given by

$$
\ell(\Phi, \Lambda, p | Y) = \sum_{i=1}^n \log \left(\sum_{k=1}^K p_k f_{(\Phi\Lambda)._{,k}}(Y_i) \right).
$$

This model can be seen as offering a connection between "model-free clustering" relying on NMF or spectral clustering as in [15, 56] and model-based clustering. Unrestricted mixture models can of course be seen as a special case by taking $H = K$ and $\Lambda = I_K$.

4

Remark 2.1. *The first example we have in mind is the mixture of multinomials that was used in [16] to model travel patterns. As our main application, this example is detailed in Subsection 2.3. Note a similarity with the Latent Dirichlet Allocation (LDA) model in [7]: LDA involves two layers of multinomials. First, a topic is a multinomial on words, then a text is described by a multinomial on topics. However, LDA does not involve clusters of similar texts. It was not designed as a clustering tool.*

Beyond multinomials, any distribution with nonnegative parameters can be used. Consider sales analysis. Assume that the owner of a supermarket observes, for each good $m \in \{1, \ldots, M\}$ and each customer $i \in \{1, \ldots, n\}$, the number of items of m bought by i during one year: $Y_{i,m}$. Put $Y_i = (Y_{i,1}, \ldots, Y_{i,M})$. We propose the model $Y_{i,m}|(Z_{i,k} = 1) \sim \mathcal{P}(\theta_{m,k})$, a Poisson distribution. The column $\theta_{\cdot,k}$ is the "standard basket" of any customer i in cluster k. But the number of goods is so huge that the estimation of standard baskets is subject to a large variance, and prevents their interpretation. In (2), the columns of Φ are representations of columns of θ in a smaller subspace. It is likely that substituable goods are gathered. This example is simply a model-based version of the NMF analysis used in [25, 29], see also [53] for an early application on the Netflix prize data. Sales analysis, customer clustering and recommender systems are indeed applications of NMF that generated a huge number of publications. More examples could include exponential or gamma mixtures in survival analysis, or Pareto and Weibull mixtures in extreme analysis.

We now discuss the adaptation of the EM algorithm to this parameter restriction.

2.2   The NMF-EM algorithm

We remind the expression of the completed likelihood

$$\ell(\Phi, \Lambda, p|Y, Z) = \sum_{i=1}^{n} \sum_{k=1}^{K} Z_{i,k} \log \left(p_k f_{(\Phi\Lambda)_{\cdot,k}}(Y_i) \right).$$

A step of the EM algorithm, given current parameters $(\Phi^{(c)}, \Lambda^{(c)}, p^{(c)})$ is as follows:

$$
\begin{aligned}
\textbf{E-step}: Q^{(c)}(\Phi, \Lambda, p) &= \mathbb{E}_{\Phi^{(c)}, \Lambda^{(c)}, p^{(c)}}[\ell(\Phi, \Lambda, p|Y, Z)|Y] \\
&= \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{E}_{\Phi^{(c)}, \Lambda^{(c)}, p^{(c)}}[Z_{i,k}|Y] \log \left(p_k f_{(\Phi\Lambda)_{\cdot,k}}(Y_i) \right) \\
\text{and } t_{i,k}^{(c)} &:= \mathbb{E}_{\Phi^{(c)}, \Lambda^{(c)}, p^{(c)}}[Z_{i,k}|Y] = \frac{p_k^{(c)} f_{(\Phi^{(c)}\Lambda^{(c)})_{\cdot,k}}(Y_i)}{\sum_{k'=1}^{K} p_{k'}^{(c)} f_{(\Phi^{(c)}\Lambda^{(c)})_{\cdot,k'}}(Y_i)}. \quad (4)
\end{aligned}
$$

$$\textbf{M-step}: (\Phi^{(c+1)}, \Lambda^{(c+1)}, p^{(c+1)}) := \underset{\Phi_{j,h}, \Lambda_{h,k} \geq 0}{\arg\max}\ Q^{(c)}(\Phi, \Lambda, p). \quad (5)$$

Obviously, the challenging step is the M-step. While we obviously have, for $k \in \{1, \ldots, K\}$,

$$p_k^{(c+1)} = \frac{\sum_{i=1}^{n} t_{i,k}^{(c)}}{\sum_{i=1}^{n} \sum_{k'=1}^{K} t_{i,k'}^{(c)}}, \quad (6)$$

5

the nonnegativity constraint on Φ and Λ makes the optimization with respect to these two matrices much harder. This is where one has to plug ideas from the NMF literature. Many options might be possible, depending on the form of $f_\vartheta(\cdot)$. The most commonly used algorithm is the so-called multiplicative update, an alternating optimization method with respect to Φ and Λ, that was proposed in the seminal papers [26, 27]. Other algorithms include ADMM [10, 50], alternating projected gradient [28], and for Bayesian approaches, Monte-Carlo methods [41] and variational approximations [1]. A numerical comparison of many algorithms can be found in [28]. In practice, the multiplicative update is efficient in many settings and is very simple to use: it does not depend on any tuning parameter such as the step size in gradient based method. So this is the method we will use from now. This method iterates a step in Φ, and a step in Λ. Each step is shown to improve the fit criterion in [27]. Note that the author claims that it also leads to convergence, but as argued in [21] the proof of this fact is actually incomplete. We explicit the multiplicative update in the case of mixture of multinomials below.

2.3   The NMF-EM algorithm for mixture of multinomials

In [16] the authors modeled a passenger temporal profile by a mixture of multinomial distribution. The time and days of smart card validations of a passenger i are recorded over a period of time (e.g. 1 month). The numbers of journeys, N_i, is not our variable of interest, and will be considered as deterministic. We obtain as a result a vector $Y_i = (Y_{i,1}, \ldots, Y_{i,M})^T \in \mathbb{R}^M$ where each coordinates represents the number of travels at a given pair time-day during the considered period. Note that of course $\sum_{k=1}^M Y_{i,k} = N_i$, let $N = \sum_{i=1}^n N_i$ be the total number of journeys. We consider a hourly grid, that is, Mon-12am, Mon-1am, etc... to Sun-11pm, with means that $M = 7 \times 24 = 168$. An example of a traveler profile is given in Figure 1.
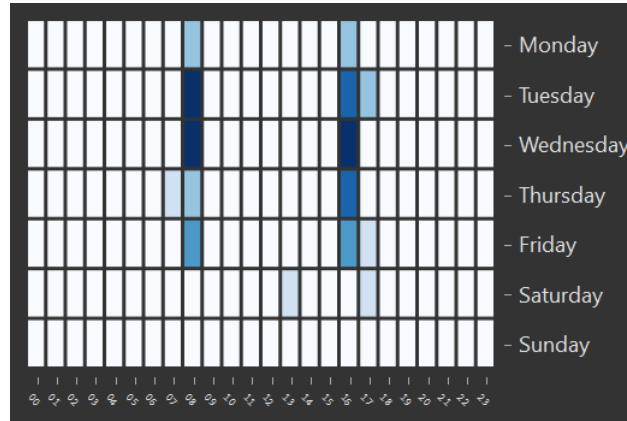

Figure 1: Temporal profile of a network user, taken from the data described in Section 4. Opacity is proportional to the number of smart-card validations.

It is natural to assume that there are clusters of passengers with rather similar profiles: for examples, employees with similar work hours or students in the same University are likely to commute at similar times. We follow the previous construction: we define the hidden cluster

6

variables, $Z_i \sim \mathcal{M}ult(p, 1)$ for some $p \in \mathcal{S}_K$. We then set

$$Y_i \big| (Z_{i,k} = 1) \sim \mathcal{M}ult(\theta_{\cdot,k}, N_i)$$

where $\theta_{\cdot,k} \in \mathcal{S}_M$ is the k-th column of an $M \times K$ matrix θ that satisfies $\theta = \Phi\Lambda$ where Φ is $M \times H$ and Λ is $H \times K$ for some $H \le M, K$. The log-likelihood is given by

$$\ell(\Phi, \Lambda, p | Y) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} p_k \left[N_i! \prod_{j=1}^{M} \frac{(\Phi\Lambda)_{j,k}^{Y_{i,j}}}{Y_{i,j}!} \right] \right\}.$$

Note that a simple way to ensure $\theta_{\cdot,k} \in \mathcal{S}_M$ is to impose similar constraints on the columns of Φ and Λ. So we define $\mathcal{M}_{M,H,K}$ as the set of all pairs (Φ, Λ) of matrices $M \times H$ and $H \times K$ respectively, with $\Phi_{\cdot,k}, \Lambda_{\cdot,j} \in \mathcal{S}_M$ for any k and j. Note that we actually have $H(M-1) + K(H-1) + K - 1$ degrees of freedom for the parameters of our model $(\Phi, \Lambda) \in \mathcal{M}_{M,H,K}$ and $p \in \mathcal{S}_K$. This knowledge is required for computing model selection criterion such as AIC (see the discussion on model selection in Subsection 2.4 below).

Let $(\hat{\Phi}, \hat{\Lambda}, \hat{p})$ denote the MLE, that is, a maximizer of $\ell(\Phi, \Lambda, p | Y)$ with respect to $(\Phi, \Lambda) \in \mathcal{M}_{M,H,K}$ and $p \in \mathcal{S}_K$. We explicit the NMF-EM algorithm to approximate $(\hat{\Phi}, \hat{\Lambda}, \hat{p})$.

From (4), values $t_{i,k}^{(c)}$ are given by

$$t_{i,k}^{(c)} = \frac{p_k^{(c)} N_i! \prod_{j=1}^{M} \frac{\left(\sum_{h=1}^{H} \Phi_{j,h}^{(c)} \Lambda_{h,k}^{(c)} \right)^{Y_{i,j}}}{Y_{i,j}!}}{\sum_{k'=1}^{K} p_{k'}^{(c)} N_i! \prod_{j=1}^{M} \frac{\left(\sum_{h=1}^{H} \Phi_{j,h}^{(c)} \Lambda_{h,k'}^{(c)} \right)^{Y_{i,j}}}{Y_{i,j}!}} = \frac{p_k^{(c)} \prod_{j=1}^{M} \left(\sum_{h=1}^{H} \Phi_{j,h}^{(c)} \Lambda_{h,k}^{(c)} \right)^{Y_{i,j}}}{\sum_{k'=1}^{K} p_{k'}^{(c)} \prod_{j=1}^{M} \left(\sum_{h=1}^{H} \Phi_{j,h}^{(c)} \Lambda_{h,k'}^{(c)} \right)^{Y_{i,j}}}.$$

We have

$$Q^{(c)}(\Phi, \Lambda, p) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{i,k}^{(c)} \log \left(p_k N_i! \prod_{j=1}^{M} \frac{\left(\sum_{h=1}^{H} \Phi_{j,h} \Lambda_{h,k} \right)^{Y_{i,j}}}{Y_{i,j}!} \right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} t_{i,k}^{(c)} \left[\log(p_k) + \log(N_i!) \right.$$

$$\left. + \sum_{j=1}^{M} \left(Y_{i,j} \log \left(\sum_{h=1}^{H} \Phi_{j,h} \Lambda_{h,k} \right) - \log(Y_{i,j}!) \right) \right].$$

As stated in (6), $p_k^{(c+1)} \propto \sum_{i=1}^{n} t_{i,k}^{(c)}$, and

$$(\Phi^{(c+1)}, \Lambda^{(c+1)}) = \arg\max_{(\Phi,\lambda) \in \mathcal{M}_{M,H,K}} \sum_{k=1}^{K} \sum_{j=1}^{M} \left(\sum_{i=1}^{n} Y_{i,j} t_{i,k}^{(c)} \right) \log \left(\sum_{h=1}^{H} \Phi_{j,h} \Lambda_{h,k} \right).$$

Put $M_{j,k}^{(c)} = \sum_{i=1}^{n} Y_{i,j} t_{i,k}^{(c)}$ for short. The previous equation becomes

$$(\Phi^{(c+1)}, \Lambda^{(c+1)}) = \arg\max_{(\Phi,\Lambda) \in \mathcal{M}_{M,H,K}} \sum_{k=1}^{K} \sum_{j=1}^{M} M_{j,k}^{(c)} \log \left(\sum_{h=1}^{H} \Phi_{j,h} \Lambda_{h,k} \right). \tag{7}$$

The maximization in (7) is equivalent to the minimization of

$$D(M^{(c)} || \Phi\Lambda) := - \sum_{k=1}^{K} \sum_{j=1}^{M} \left\{ M_{j,k}^{(c)} \log \left(\sum_{h=1}^{H} \Phi_{j,h} \Lambda_{h,k} \right) - \sum_{h=1}^{H} \Phi_{j,h} \Lambda_{h,k} \right\}. \tag{8}$$

Indeed, for $(\Phi, \Lambda) \in \mathcal{M}_{M,H,K}$ we have $\sum_{k=1}^{K} \sum_{j=1}^{M} \sum_{h=1}^{H} \Phi_{j,h} \Lambda_{h,k} = \sum_{j=1}^{M} \sum_{k=1}^{K} (\Phi\Lambda)_{j,k} = \sum_{j=1}^{M} 1 = M$ that does not depend on (Φ, Λ). The multiplicative algorithm in [27] was actually introduced to minimize $D(M^{(c)}||\Phi\Lambda)$. So we just use the update steps of [27] (steps 9 and 10 in Algorithm 1 below) followed by a renormalization of the matrices Φ and Λ in order to ensure that the columns remain in the parameter space (steps 10 and 12). This completes the derivation of the NMF-EM algorithm for mixture of multinomials: Algorithm 1 page 8. We implemented this algorithm for the R software [24], the package `nmfem` can be found on the CRAN repository.

Algorithm 1 NMF-EM

1: Fix $\epsilon > 0$. Choose arbitrary $\Phi^{(0)}$, $\Lambda^{(0)}$ and $p^{(0)}$; $c := 0$, CRIT $:= \infty$.

2: **while** $|\ell(\Phi^{(c)}, \Lambda^{(c)}, p^{(c)}) - \text{CRIT}| > \epsilon$ **do**

3:     CRIT $:= \ell(\Phi^{(c)}, \Lambda^{(c)}, p^{(c)})$.

4:     For all $i \in \{1, \dots, n\}$ and $k \in \{1, \dots, K\}$,

$$t_{i,k}^{(c)} := \frac{p_k^{(c)} \prod_{j=1}^{M} \left(\sum_{h=1}^{H} \Phi_{j,h}^{(c)} \Lambda_{h,k}^{(c)} \right)^{Y_{i,j}}}{\sum_{k'=1}^{K} p_{k'}^{(c)} \prod_{j=1}^{M} \left(\sum_{h=1}^{H} \Phi_{j,h}^{(c)} \Lambda_{h,k'}^{(c)} \right)^{Y_{i,j}}} \text{ and } p_k^{(c+1)} =: \frac{\sum_{i=1}^{n} t_{i,k}^{(c)}}{\sum_{i=1}^{n} \sum_{k'=1}^{K} t_{i,k'}^{(c)}}.$$

5:     $\forall j, k \quad M_{j,k}^{(c)} = \sum_{i=1}^{n} Y_{i,j} t_{i,k}^{(c)}$.

6:     Initialization of Φ and Λ (arbitrarily), $q := \infty$.

7:     **while** $|Q^{(c)}(\Phi, \Lambda, p^{(c+1)}) - q| > \epsilon$ **do**

8:         $q := Q^{(c)}(\Phi, \Lambda, p^{(c+1)})$.

9:         $\forall h, k \quad \Lambda_{h,k} \leftarrow \Lambda_{h,k} \frac{\sum_j \Phi_{j,h} M_{j,k}^{(c)} / (\Phi\Lambda)_{j,k}}{\sum_j \Phi_{j,h}}$

10:       $\forall h, k \quad \Lambda_{h,k} \leftarrow \frac{\Lambda_{h,k}}{\sum_{k'} \Lambda_{h,k'}}$

11:       $\forall j, h \quad \Phi_{j,h} \leftarrow \Phi_{j,h} \frac{\sum_k \Lambda_{h,k} M_{j,k}^{(c)} / (\Phi\Lambda)_{j,k}}{\sum_k \Lambda_{h,k}}$

12:       $\forall j, h \quad \Phi_{j,h} \leftarrow \frac{\Phi_{j,h}}{\sum_{h'} \Phi_{j,h'}}$

13:     **end while**

14:     $(\Phi^{(c+1)}, \Lambda^{(c+1)}) := (\Phi, \Lambda)$.

15:     $c := c + 1$.

16: **end while**

2.4   Discussion on the choice of H and K

The choice of K is not a straightforward issue in mixture models. *A fortiori* the choice of the pair (H, K) is not easier.

From the likelihood and the degrees of freedom above we can derive the AIC and BIC criteron

$$\text{AIC} = \ell(\hat{\Phi}, \hat{\Lambda}, \hat{p}|Y) - \frac{H(M-1) + K(H-1) + K - 1}{2}$$

$$\text{BIC} = \ell(\hat{\Phi}, \hat{\Lambda}, \hat{p}|Y) - \frac{[H(M-1) + K(H-1) + K - 1]\log(N)}{2}$$

that are widely used in practice. Among the papers mentioned above, BIC is used for choosing the

number of clusters of users in [9]. However, the consistency of AIC and BIC depend on conditions that might not be satisfied in mixture models. Other criteria more suitable for mixtures were investigated, like NEC and variants [5]. The slope heuristic [3] is known to give nice results in practice, and can also be show to be consistent in some settings [2]. It is actually used in [16] for mixtures of multinomials.

An important point is that our criterion should actually depend on the objective we have in mind. In regular models, AIC finds the optimal balance between bias and variance, while BIC identifies the true model, when there is one. These two objectives are usually not compatible [55]. In our collaboration with Transdev, interpretability of the results was actually one of the main objectives. We will use the slope heuristic in what follows.

3   Simulation study

In this section, we illustrate the dimension-reduction effect of NMF-EM on synthetic data. As our main interest is here clustering, we will compare the "pairwise misclassification rate" of NMF-EM with the one of EM and k-means algorithms – that is, the proportion of pairs (i, j) of individuals that are either assigned to the same component by the algorithm while they were actually generated from different components, or assigned to different components while they were simulated from the same.

The experimental setting is as follow: the dimension is $m = 100$, for each experiment we generate H_0 words in $\mathbb{R}^m$ from a uniform distribution and then $K = 10$ parameters $\theta_{\cdot,1}, \ldots, \theta_{\cdot,K}$ as linear combinations of these H_0 words - the coefficients of each parameters are independently drawn from a Dirichlet distribution $\mathcal{D}(\alpha, \ldots, \alpha)$. We finally draw $n = 1500$ individuals from the corresponding mixture of multinomials with uniform weights.

We compare NMF-EM with $H = 4$, EM (without reduction of dimension) and k-means in various settings: in the case $H_0 = 4$, where the dimension reduction in NMF-EM is actually correct, and $H_0 = 8$ - this case is less favourable to NMF-EM with $H = 4$ as it reduces too much the dimension... We also use different values for α, leading to different shapes for the set of parameters $\{\theta_{\cdot,1}, \ldots, \theta_{\cdot,K}\}$. The results are in Tables 1 and 2.

So, when the intrinsic dimension is small enough, NMF-EM really improves the clustering ability of EM. In any case, our main claim is that it leads to easily interpretable clusters, a fact that will be illustrated in the next section.

4   Application to ticketing data

4.1   Description of the data

The data used in our study are the validations made during the month of September 2015 on one Transdev network in a medium size city. Ticketing data are the information obtained at each transaction made by a smart card on a validator system. For privacy reasons it is not possible to

Table 1: Pairwise misclassification rate of the algorithms on simulated data when $H_0 = 4$ ($m = 100$, $n = 1500$, $N = 150$, $K = 10$).

	α = .01	α = .1	α = .2	α = .3	α = .4	α = .5	α = .6
NMF-EM	9.5%	6.3%	**4.7%**	**5.0%**	**4.9%**	**5.3%**	**5.9%**
EM	8.4%	6.8%	5.0%	5.7%	5.6%	5.9%	6.7%
k-means	**6.4%**	**5.9%**	5.3%	5.5%	5.6%	6.0%	6.2%
	α = .7	α = .8	α = .9	α = 1.0	α = 1.1	α = 1.2	α = 1.3
NMF-EM	**6.5%**	**6.7%**	**6.7%**	7.6%	7.3%	7.5%	8.8%
EM	7.2%	7.3%	7.0%	7.7%	8.1%	8.0%	8.9%
k-means	6.6%	6.7%	6.8%	**7.0%**	**7.2%**	**7.1%**	**7.5%**

Table 2: Pairwise misclassification rate of the algorithms on simulated data when $H_0 = 8$ ($m = 100$, $n = 1500$, $N = 150$, $K = 12$).

	α = .01	α = .1	α = .2	α = .3	α = .4	α = .5	α = .6
NMF-EM	5.2%	4.5%	5.8%	5.8%	6.5%	6.9%	8.1%
EM	4.3%	3.1%	**3.1%**	**3.8%**	5.0%	6.1%	6.1%
k-means	**3.8%**	**3.1%**	3.4%	4.0%	**4.8%**	**5.5%**	**5.6%**
	α = .7	α = .8	α = .9	α = 1.0	α = 1.1	α = 1.2	α = 1.3
NMF-EM	8.2%	9.1%	10.0%	10.5%	10.3%	11.3%	11.5%
EM	6.4%	7.2%	7.5%	7.5%	8.3%	8.6%	8.5%
k-means	**5.8%**	**6.3%**	**6.3%**	**6.5%**	**6.8%**	**7.0%**	**6.9%**

connect each validation to the user who made it. The feature that allows us to realize our study and create temporal profiles is a card number which is encrypted, and re-initialized every three months. It is thus impossible to follow the long-term behaviour of a user. This is the reason why we focus on a one month period. This period (September) have been chosen because it has no vacation nor bank holiday. During September 2015, more than $4,000,000$ check-ins have been made on the network by $232,430$ passengers.

The data are agregated so that for each traveler, for each day of the week (Monday to Sunday) and each hour (00 to 23), we have the number of validation during the studied period. A passenger profile is thus defined by $24 * 7 = 168$ features. Figure 1 page 6 already provided an example of a temporal profile of one of the users. This traveler uses mainly the network at 8 a.m and 4 p.m.

Remark 4.1. *We used the same strategy to create stations profiles: for each station, for each day of the week and each hour of the day, we know the number of validations that occured at this station during the study period. In Figure 2, we show the temporal profile of the station "Palais de Justice" (courthouse), a tramway station in the city center. This station has travelers all day long, but knows an attendance peak every day from 4 to 6 p.m. The results of this analysis are*

Figure 2: Temporal profile of station "Palais de Justice".

provided in the supplementary material.

In order to avoid users who would not use their smart card enough to exhibit a clear pattern, data have been cleaned. We define a "regular card holder" as a card holder who

- travelled on at least four days during September 2015 (so in particular we have $N_i \geq 4$);

- made their first boarding after 4 a.m each day at the same station 50% of the time.

We only kept regular card holders for our analysis. After this cleaning step, we end up with $72,359$ profiles of passengers, which represent a bit more than $3,000,000$ check-ins – that means 31% of passengers represent 75% of check-ins. We also have 475 stations profiles. These data are provided in the `nmfem` package.

4.2 Passenger profile clustering

We first focus on passengers profiles clustering. This allows us to create groups of people that have similar temporal habits. The method used to create these clusters is the NMF-EM algorithm from Subsection 2.3.

To choose the parameters H and K, we begin with the analysis of the log-likelihood of our model when $H = K$ for $K = 2 \ldots 30$. Note that the estimation of the model in this case can be made by the usual EM algorithm for multinomial mixture model. Figure 3 shows the evolution of the log-likelihood as a function of K. This function clearly exhibits a linear behavior when

Figure 3: The log-likelihood as a function of K under $H = K$.

$K \geq 10$. Thus, the slope heuristic suggests considering $K = 10$.

Now keeping $K = 10$ fixed, we chose the value of H in the same way. First, we plot the log-likelihood as a function H in Figure 4. By using again the slope heuristic method, we choose

Figure 4: The log-likelihood as a function of $H \in \{2, \ldots, K\}$ under $K = 10$.

$H = 5$.

The $H = 5$ words and the $K = 10$ clusters are represented in Figure 5 and in Figure 7 respectively. Remember that each cluster can be decomposed as a convex combination of words, some of them might have a null weight. For example, Figure 6 shows how the parameter of Cluster 5 can be written as a convex combination of words 4 and 2.

Figure 5: Words obtained by NMF-EM on users data with $K = 10$ and $H = 5$.

The interpretation of the words is direct:

1. Word 1: travels between 6 a.m and 7 a.m.

2. Word 2: diffuse component during off-peak periods (i.e. from 9 a.m to 4 p.m).

3. Word 3: travels at school hours. Indeed it is composed of travel between 7 and 8 a.m and between 4 and 5 p.m, except on Wednesdays, when the afternoon travel is replaced by one at noon.

4. Word 4: travels between 8 and 9 a.m.

5. Word 5: late afternoon peak, from 5 to 7 p.m, and Wednesdays and Saturdays afternoon.

We now attempt an interpretation of the clusters:

Figure 6: Decomposition of cluster 5 from words 4 and 2.

1. Clusters 1, 3, 4 and 6 present high travel probabilities in the morning and in the afternoon except Wednesdays where the afternoon travel is replaced by a higher probability of travel around noon. These four clusters are typical of French schools and high-schools hours. The main differences are:

   (a) Cluster 1: travels at 7 a.m and around 4 or 5 p.m.

   (b) Cluster 3: travel a bit more at 8 a.m.

   (c) Cluster 4: travelers are less susceptible to travel after 5 p.m.

   (d) Cluster 6: travels at 6 and 7 a.m.

2. Cluster 5: travels at 8 a.m and at 4 or 5 p.m.

3. Cluster 7: travels mainly at 6 a.m.

4. Cluster 9: travels at 8 a.m and at 5 p.m.

5. Clusters 2, 8 and 10: diffuse travel habits.

   (a) Cluster 2: travels Mondays to Saturdays from 7 a.m to 7 p.m with highest probabilities at 8 a.m and 5 p.m Mondays to Fridays.

   (b) Cluster 8: diffuse travels Mondays to Saturdays from 9 a.m to 7 p.m.

   (c) Cluster 10: travels Mondays to Fridays from 9 a.m to 4 p.m.

Figure 7: Clusters obtained by NMF-EM on users data with $K = 10$ and $H = 5$.

As a conclusion, NMF-EM provides clusters of users that are easily interpretable. In the supplementary material, we show how users profiles are related to demographic and economic variables.

5 Conclusion

We provided a new approach for dimension reduction that can be compared to MFA in non-Gaussian mixture models. This approach is based on NMF, an extremely popular data mining algorithm. We adapted the EM algorithm to this setting. This new algorithm, NMF-EM, is implemented in a package for R in the case of mixtures of multinomials. Results on simulated and real data are promising. In addition to a theoretical study of algorithm, future work should include an application to mixture of other distribution with nonnegative parameters like the Poisson distribution.

Acknowledgements. We would like to thank the anonymous Referees and the Associate Editor for their constructive comments and suggestions. We also thank Denis COUTROT and Nadir MEZIANI from Transdev for their support and comments on previous versions of this work.

<div align="center">

SUPPLEMENTARY MATERIAL

</div>

Appendix Contains the analysis of the clusters of users in Section 4. We also apply NMF-EM on stations profile, and fully analyze another network (in the Netherlands).

References

[1] P. Alquier and B. Guedj. An oracle inequality for quasi-Bayesian non-negative matrix factorization. *Mathematical Methods of Statistics*, 26(1):55–67, 2017.

[2] S. Arlot and P. Massart. Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10(Feb):245–279, 2009.

[3] M. C. Baudry, J.-P. and B. Michel. Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470, 2012.

[4] T. Benaglia, D. Chauveau, D. R. Hunter, and D. Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009.

[5] C. Biernacki, G. Celeux, and G. Govaert. An improvement of the NEC criterion for assessing the number of clusters in a mixture model. *Pattern Recognition Letters*, 20(3):267–272, 1999.

[6] C. Bishop. Pattern recognition and machine learning (information science and statistics), 1st edn. 2006. corr. 2nd printing edn, 2007.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[8] C. Bouveyron and C. Brunet-Saumard. Model-based clustering of high-dimensional data: a review. *Computational Statistics and Data Analysis*, 71:52–78, 2014.

[9] C. Bouveyron, E. Côme, and J. Jacques. The discriminative functional mixture model for a comparative analysis of bike sharing systems. *The Annals of Applied Statistics*, 9(4):1726–1760, 2015.

[10] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[11] L. Carel and P. Alquier. Non-negative matrix factorization as a pre-processing tool for travelers temporal profiles clustering. In M. Verleysen, editor, *Proceedings of the 25th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, pages 417–422. i6doc.com, 2017.

[12] G. Celeux, S. Frühwirth-Schnatter, and C. P. E. Robert. *Handbook of Mixture Analysis*. CRC Press, 2018.

[13] G. Celeux, C. Maugis-Rabusseau, and M. Sedki. Variable selection in model-based clustering and discriminant analysis with a regularization approach. To appear in Advances in Data Analysis and Classification, 2018.

[14] E. Côme and L. Oukhellou. Model-based count series clustering for bike sharing system usage mining: a case study with the Vélib' system of Paris. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):39, 2014.

[15] C. Ding, X. He, and H. D. Simon. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM International Conference on Data Mining*, pages 606–610. SIAM, 2005.

[16] M. K. El Mahrsi, E. Côme, J. Baro, and L. Oukhellou. Understanding passenger patterns in public transit through smart card and socioeconomic data: a case study in Rennes, France. In *ACM SIGKDD Workshop on Urban Computing*, 2014.

[17] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis. *Neural computation*, 21(3):793–830, 2009.

[18] M. Fop and T. B. Murphy. Variable selection methods for model based clustering. *arXiv preprint arXiv:1707.00306*, 2017.

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

[20] Z. Ghahramani and G. E. Hinton. The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[21] E. F. Gonzalez and Y. Zhang. Accelerating the Lee-Seung algorithm for non-negative matrix factorization. *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*, 2005.

[22] B. Grün. Model-based clustering. In G. Celeux, S. Frühwirth-Schnatter, and C. P. Robert, editors, *Handbook of Mixture Analysis*, pages 155–188. CRC Press, 2018.

[23] R. Hamon, P. Borgnat, C. Févotte, P. Flandrin, and C. Robardet. Factorisation de réseaux temporels: étude des rythmes hebdomadaires du système Vélo'v. In *Colloque GRETSI 2015*, 2015.

[24] R. Ihaka and R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5.

[25] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[26] D. Lee and S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[27] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[28] C.-J. Lin. Projected gradient methods for non-negative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[29] X. Luo, M. Zhou, Y. Xia, and Q. Zhu. An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems. *IEEE Transactions on Industrial Informatics*, 10(2):1273–1284, 2014.

[30] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709, 2009.

[31] C. Maugis, G. Celeux, and M.-L. Martin-Magniette. Variable selection in model-based clustering: a general variable role modeling. *Computational Statistics & Data Analysis*, 53(11):3872–3882, 2009.

[32] G. McLachlan and D. Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[33] G. J. McLachlan, D. Peel, and R. W. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics & Data Analysis*, 41(3-4):379–388, 2003.

[34] P. D. McNicholas. *Mixture model-based classification*. CRC Press, 2016.

[35] P. D. McNicholas. Model-based clustering. *Journal of Classification*, 33(3):331–373, 2016.

[36] P. D. McNicholas and T. B. Murphy. Parsimonious gaussian mixture models. *Statistics and Computing*, 18(3):285–296, 2008.

[37] J. Mei, Y. De Castro, Y. Goude, and G. Hébrail. Recovering multiple nonnegative time series from a few temporal aggregates. In *34th International Conference on Machine Learning (ICML)*, 2017.

[38] A. Montanari and C. Viroli. Heteroscedastic factor mixture analysis. *Statistical Modelling*, 10(4):441–460, 2010.

[39] C. Morency, M. Trepanier, and B. Agard. Measuring transit use variability with smart-card data. *Transport Policy*, 14(3):193–203, 2007.

[40] K. Murphy, I. C. Gormley, and C. Viroli. Infinite mixtures of infinite factor analysers: nonparametric model-based clustering via latent gaussian models. *arXiv preprint arXiv:1701.07010*, 2017.

[41] J. W. Paisley, D. M. Blei, and M. I. Jordan. Bayesian nonnegative matrix factorization with stochastic variational inference., 2014.

[42] M.-P. Pelletier, M. Trépanier, and C. Morency. *Smart card data in public transit planning: a review*. CIRRELT, 2009.

[43] C. Peng, X. Jin, K.-C. Wong, M. Shi, and P. Liò. Collective human mobility pattern from taxi trips in urban area. *PloS one*, 7(4):e34487, 2012.

[44] M. Poussevin, E. Tonnelier, N. Baskiotis, V. Guigue, and P. Gallinari. Mining ticketing logs for usage characterization with nonnegative matrix factorization. In *International Workshop on Modeling Social Media*, pages 147–164. Springer, 2014.

[45] A. E. Raftery and N. Dean. Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178, 2006.

[46] A. N. Randriamanamihaga, E. Côme, L. Oukhellou, and G. Govaert. Clustering the Vélib' origin-destinations flows by means of poisson mixture models. In *ESANN*, 2013.

[47] L. Scrucca, M. Fop, and A. E. Murphy, T. B.and Raftery. mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R journal*, 8(1):289, 2016.

[48] F. Shahnaz, M. Berry, P. Pauca, and R. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

[49] D. Steinley and M. J. Brusco. Selection of variables in cluster analysis: an empirical comparison of eight procedures. *Psychometrika*, 73(1):125–144, 2008.

[50] D. Sun and C. Févotte. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6201–6205. IEEE, 2014.
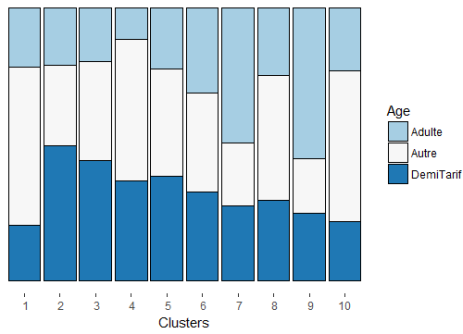
[51] E. Tonnelier, N. Baskiotis, V. Guigue, and P. Gallinari. Anomaly detection in smart card logs and distant evaluation with twitter: a robust framework. *To appear in Neurocomputing*, 2018.

[52] J. H. Wolfe. Object cluster analysis of social areas. MasterâĂŹs thesis, University of California, 1963.

[53] M. Wu. Collaborative filtering via ensembles of matrix factorizations. In *Proceedings of KDD Cup and Workshop*, volume 2007, 2007.

[54] W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

[55] Y. Yang. Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, 92(4):937–950, 2005.

[56] Z. Yang, J. Corander, and E. Oja. Low-rank doubly stochastic matrix decomposition for cluster analysis. *Journal of Machine Learning Research*, 17(187):1–25, 2016.

[57] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(3):38, 2014.

Appendix

Analysis of the clusters of users

As written above, we have no personal information in our data. Therefore, we are not able to describe individually the users in each cluster. However, for each transaction made, we have the encrypted card number and the transport ticket used. So we can recover for each card the most used transport ticket during the period. This provides interesting information as some schemes are associated to age ranges (Young, Senior...) and to time periods (Unit, Annual or Monthly Subscription). Let us now provide the description of each cluster in terms of age ranges (Figures 3a to 3c in Table 3).

(a) Adults ("Adult") and Reduced tickets ("DemiTarif")

(b) Children ("Enfant") and Youngs ("Jeune")

(c) Free travelers ("Gratuit") and Elderly ("Senior")

Table 3: Age range analysis of the clusters

Adults are more present in clusters 7 and 9, that are clusters with check-ins mostly in the morning. People benefiting from half-price are present in every cluster but with highest rates in clusters 2, 3, 4 and 5. Children (4 to 6) are not very present on the network, but they are more represented in clusters 1, 5 and 9. Young travelers (6 to 25) are more present in clusters 1 and 4. These clusters correspond to scholar time slot. In clusters 8 and 10 there are large rate of seniors and free travelers. As these clusters have profiles of diffuse travels during the week and as free travelers are unemployed or low salaries people, these regroupments make sense.

Figure 8 shows the repartition of transport ticket type through clusters. Unit products are

Figure 8: Transportation ticket type analysis of the clusters.

more used in clusters 8 and 10 that are clusters with a lots of seniors and free travelers. As they don't have obligations, they likely use unit products for occasional trips. Clusters 1, 3, 4 and 9, that have mostly scholar profiles althought have a large majority of annual subscripters. A possible interpretation is that schoolchildren and students are public transportation captives, and have to use the network in order to go to class every day. Thus, buying an annual pass is more advantageous than buying any other product type.

As described in Subsection 4.1, we kept only users whose first trip of the day is made at the same station at least 50% of the study time. That main "morning station" is thus called the "home station" as it gives us an estimation of the residence place of users. In Tables 4 and 5, we can observe the shares of clusters by home stations. It shows the share of travelers identified as belonging to every cluster leaving near each station.

We note that:

1. Cluster 1: travelers are over represented at peripheral stations.

2. Cluster 2: no particular pattern observed.

3. Cluster 3: no particular pattern observed.

4. Cluster 4: few stations show over representation of cluster 4.

5. Cluster 5: over representation of the cluster at two stations in the north.

6. Cluster 6: no particular pattern observed.

7. Cluster 7: One station is 100% represented by cluster 7. As only one user is assigned to this station, no particular pattern is observed.

8. Cluster 8: the cluster is over represented at one station in the city center and at another further.

9. Cluster 9: cluster 9 is over represented in few stations in the center.

10. Cluster 10: cluster is over represented in poorest neighborhoods of the city.

(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

(f) Cluster 6

Table 4: Share of clusters per home station — Clusters 1 to 6

Stations profile clustering

Clustering the different stations of the network would allow us to better know the different type of stations, and to group them by temporal similarity. As we have very few number of stations (475), it is not safe to process as described above for the users clustering. Indeed, a K larger than 6 or 7 leads to very small clusters. In place we fixed H and K *a priori* to 3 and 5 respectively.

The 3 words obtained are the ones in Figure 9. The first time component is described by check-ins at 7 and 8 a.m. We will call it the "morning component". The second time component shows check-ins at 4 and 5 p.m on Mondays, Tuesdays, Thursdays and Fridays and check-ins at 12 p.m on Wednesdays. We will name it the "end of school component". The third component

(a) Cluster 7

(b) Cluster 8

(c) Cluster 9

(d) Cluster 10

Table 5: Share of clusters per home station — Clusters 7 to 10

Figure 9: Words obtained by NMF-EM on stations data with $K = 5$ and $H = 3$.

shows check-ins at 6 p.m, during Wednesdays afternoons, during Saturdays and off-peaks periods. This component will be called the "off-peak component".

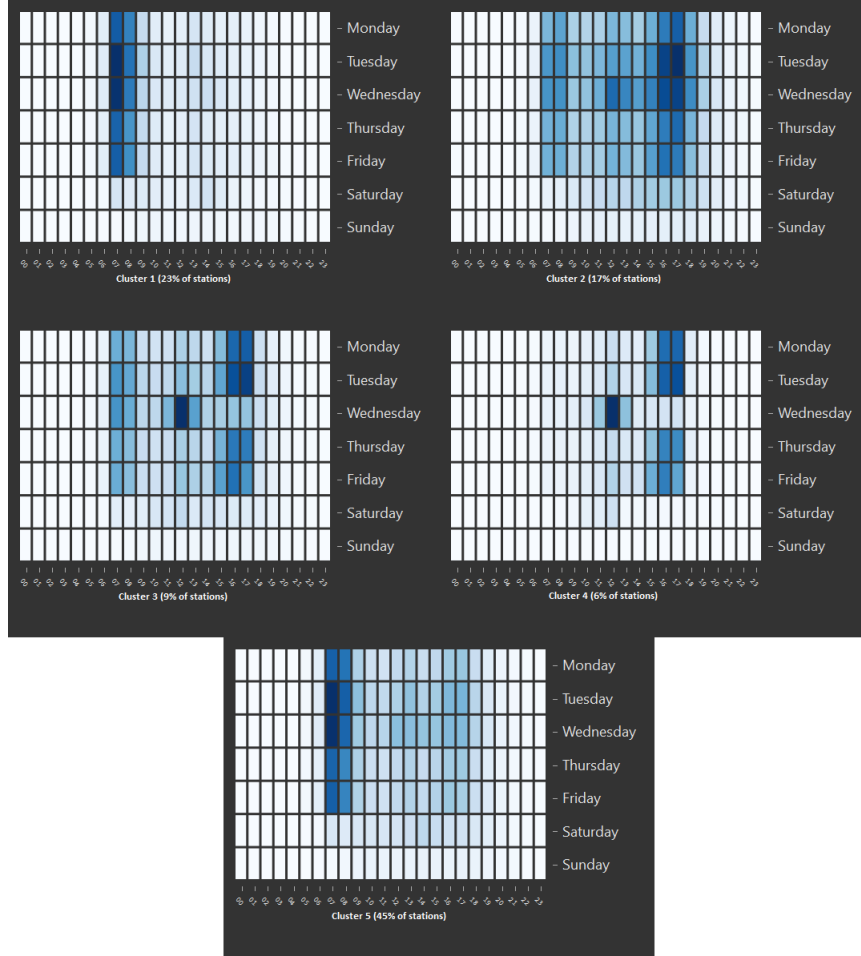Figure 10 shows the 5 clusters. Stations in cluster 1 are stations where there are check-ins

Figure 10: Clusters obtained by NMF-EM on stations data with $K = 5$ and $H = 3$.

only in the morning at 7 or 8 a.m. These stations are likely in residential areas. In cluster 2, the stations have check-ins all day long, but with highest probabilities during peaks. Stations in cluster 3 have check-ins in the morning and at the end of school. They are likely to be near schools in residentials area. Stations in cluster 4 have check-ins only at end of school times. Thus, these stations are probably near schools. Finally, stations in cluster 5 are pretty similar than the ones in cluster 1: a large majority of check-ins are made in the morning (7 or 8 p.m). The only difference is that it is more likely to have check-ins during the rest of the day in cluster 5 than in cluster 1.

Thanks to the French National Institute of Statistics and Economis Studies (INSEE), there are open data permitting us to introduce contextual information. Firstly, a database containing socioeconomic data on a grid of $200m \times 200m$ is available. We used two indicator of it: the number of inhabitants and the percentage of households living in collective housing per tiles. Secondly, we used a database referencing and geolocating every french company or administration. In this way, we were able to know the number of employees per tile. By clustering the tiles in the study area, we obtained different group of areas that will allow us to lead the study on stations more finely. Table 6 contains the description of the mean tile by cluster.
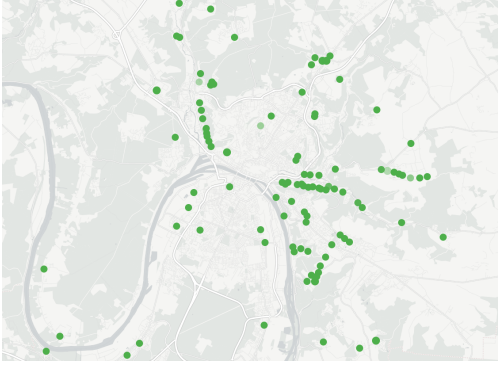
Table 6: Description of tiles clusters

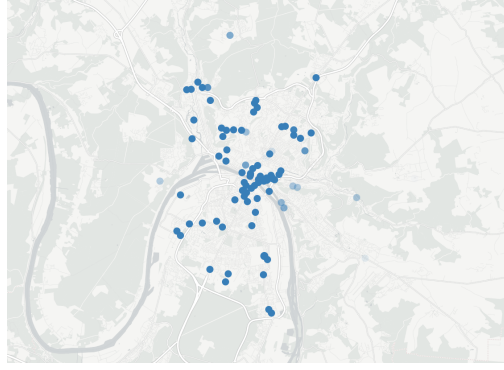| Tiles cluster	Inhabitants	Percentage of collective housing	Employees
1	223.75	57.73	824.43
2	162.03	30.08	40.32
3	268.74	58.66	2758.97
4	114.50	98.98	11576.50

As tiles contained in cluster 1 and 2, are those with the least number of employees, they can be described as residential areas. Moreover, the percentage of collective housing allows to distinguish them. Indeed, cluster 1 have more households living in collective housing than cluster 3. That is why we will refer as tiles from cluster 1 as residential areas in collective housing and as residential areas in individual housing for tiles from cluster 2. Since the number of inhabitants and of employees are high, tiles from cluster 3 will be refered as mixed areas. Finally, as the number of employees in cluster 4 is very large, we will refer these tiles as business areas.

The figures in Table 7 show the geographical repartition of the five clusters. In Figure 7a, we oserve the stations contained in cluster 1. This cluster groups stations that have check-ins only in the morning. On the figure, we observe that these stations are distant from the city center and are mainly located in residential areas. Figure 7b shows stations of cluster 2, that have check-ins all day long with stronger attendance during peak-periods. These stations are mainly located in the city center. Figures 7c and 7d look alike. Indeed, clusters 3 and 4 have the "end of school" component and the points on the map are close to educational establishment. Figure 7e shows stations from cluster 5. These stations have check-ins all day long but most are made in
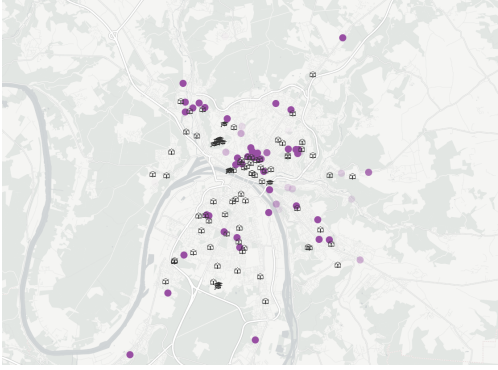
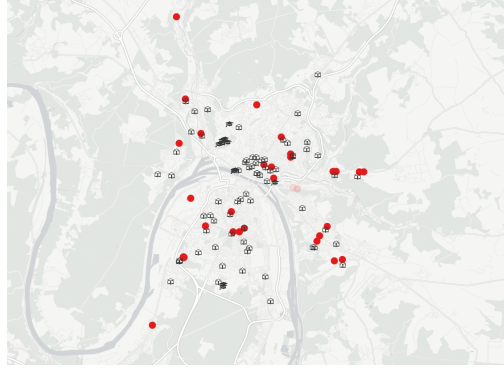the morning. By looking at the map, we cannot notice any significant pattern.
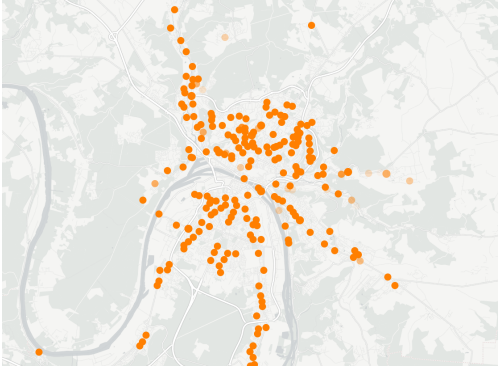

(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

(d) Cluster 4

(e) Cluster 5

Table 7: Map of the stations — opacity of the points are proportional to the adequacy between the stations and the clusters.

Passengers profile clustering on another network

To ensure efficiency of the algorithm, we applied it on another network located in the Netherlands. By applying the same model selection method as in Section 4.2, we obtained the optimal values of $K = 10$ and $H = 7$. Figures 11 and 12 contain respectively the profiles of the words and clusters obtained.
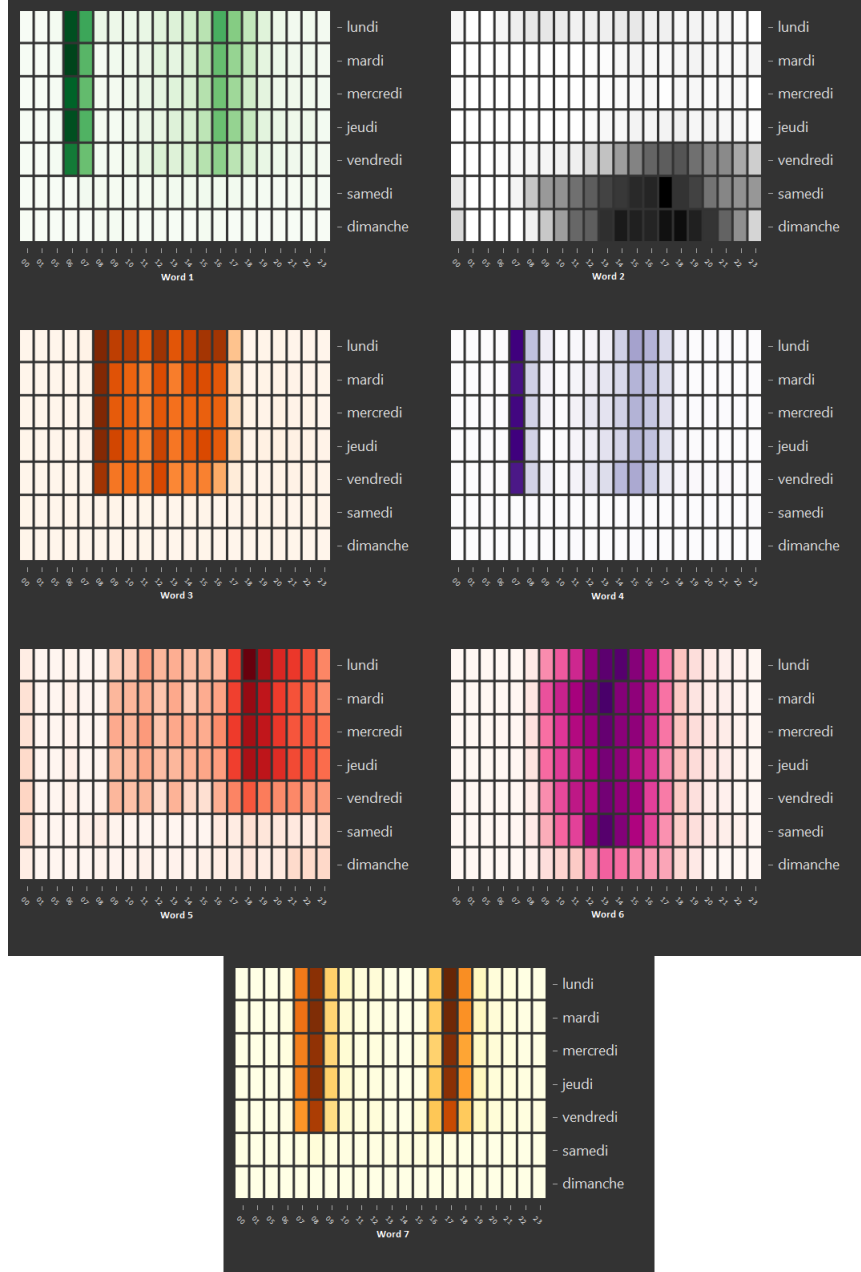
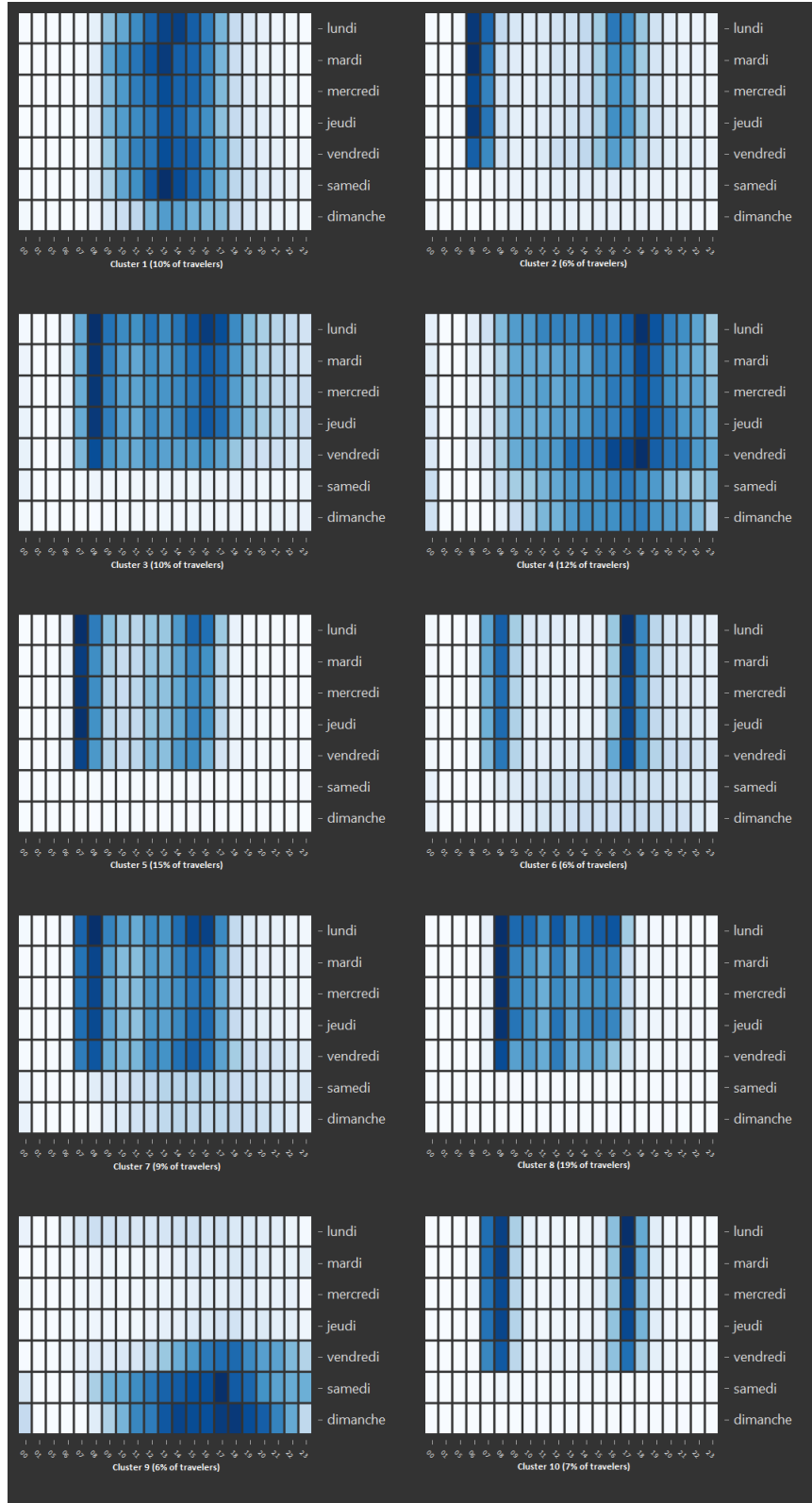Figure 11: Words obtained by NMF-EM on users data with $K = 10$ and $H = 7$.

Figure 12: Clusters obtained by NMF-EM on users data with $K = 10$ and $H = 7$.

The interpretation of the words is:

1. Word 1: travels at 6 or 7 a.m and slightly around 4 p.m during the week.

2. Word 2: travels during the week-end.

3. Word 3: diffuse travel habits from 8 a.m to 4 p.m Mondays to Fridays.

4. Word 4: travels at 7a.m on weekdays.

5. Word 5: diffuse habits with highest probabilities from 5 p.m to 12 a.m during the week.

6. Word 6: diffuse habits from 9 a.m to 5 p.m with highest probability at 1 p.m Mondays to Saturdays.

7. Word 7: travels at 8 a.m and 5 p.m.

We can interpret the cluster as follows:

1. Cluster 1: diffuse habits from 9 a.m to 5 p.m with highest probability at 1 p.m Mondays to Saturdays.

2. Cluster 2: travels at 6 or 7 a.m and at 4 or 5 p.m during the week.

3. Cluster 3: diffuse habits from 7 a.m to 6 p.m on weekdays.

4. Cluster 4: diffuse travel habits from 9 a.m to 11 p.m.

5. Cluster 5: travels at 7 or 8 a.m diffuse habits during the afternoon.

6. Cluster 6: travels at 8 a.m and 5 p.m.

7. Cluster 7: diffuse travel habits from 7 a.m to 5 p.m Mondays to Fridays.

8. Cluster 8: diffuse habits from 8 a.m to 4 p.m during the week.

9. Cluster 9: travels during the week-end.

10. Cluster 10: travels at 7 or 8 a.m and around 4 p.m.