Evolutionary clustering for categorical data using parametric links among multinomial mixture models

Md Abul Hasnat, Julien Velcin, Stephane Bonnevay, Julien Jacques

Evolutionary clustering for categorical data using parametric links among multinomial mixture models

Md. Abul Hasnat[a], Julien Velcin[a], Stephane Bonnevay[b], Julien Jacques[a]

[a] *Université de Lyon, Lumière, ERIC*
[b] *Université de Lyon, Claude Bernard, ERIC*

Abstract

In this paper, we propose a novel evolutionary clustering method for temporal categorical data based on parametric links among multinomial mixture models. Besides clustering, our main goal is to interpret the evolutions of clusters over time. To this aim, first we propose the formulation of a generalized model that establishes parametric links among two multinomial mixture. Afterward, different parametric sub-models are defined in order to model typical evolutions of the clustering structure. Model selection criteria allow to select the best sub-models and thus to guess the clustering evolution. For the experiments, first we evaluate the proposed method with synthetic temporal data. Next, we apply it to analyze the annotated social media data. Results show that the proposed method is better than the state-of-the-art based on the common evaluation metrics. Additionally, it can provide interpretation about the temporal evolution of the clusters.

Key words: evolutionary clustering, multinomial distribution, mixture model, model-based clustering, Twitter data

1. Introduction

In the recent years, the social media plays a significant role in many aspects of our daily activity. There exist numerous popular social media such as Twitter or Facebook, where the users (people) often provide their opinions about particular entity, e.g., persons (politician, actor), products consumed in the daily life, etc. A common method to analyze such data is to use a clustering method that naturally groups the users/opinions, and then investigate each group independently. An important property of these data is that they may change *over time* due to changes of the attributes, and appearance/disappearance of users. Moreover, users may change their opinion about the targeted entity.

An ordinary clustering method is unlikely to adapt with such temporal dynamics of the data, as it does not consider any relevant information such as history and temporal effects. The notion of evolutionary clustering [11, 37, 12, 39] appears in such situations, where the method should be specialized in clustering temporal data by taking care of the historic information and current data altogether. Numerous methods exist, which address these issues appropriately and cluster temporal data. These methods are based on different strategies, such as spectral clustering [12, 37] and probabilistic generative model [9, 39, 23]. However, it remains an important issue - how to interpret the evolution of the clusters. In this research, we are motivated by this issue and propose a novel method based on the multinomial mixture model [8] to cluster the temporal data as well as interpret the evolution of the clusters through some prior belief. Therefore, we propose a novel method which simultaneously performs evolutionary clustering and interpreting the evolution.

Multinomial Mixture (MM) model based clustering strategy is a popular method for clustering discrete data [27, 34, 18, 1]. Most recently, it has been exploited to perform evolutionary clustering [23]. In this

Email addresses: mhasnat@gmail.com (Md. Abul Hasnat), julien.velcin@univ-lyon2.fr (Julien Velcin), stephane.bonnevay@univ-lyon1.fr (Stephane Bonnevay), julien.jacques@univ-lyon2.fr (Julien Jacques)

research, we consider MM as the core model for the data and propose an evolutionary clustering method by deriving appropriate link between the parameters of MM at different time.

Parametric link among probability distributions has been used in the context of transfer learning [4, 21, 3], where the goal is to adapt a clustering model from a source population to a target one. In the context of continuous features, [4] proposed a parametric link between the Normal distributions. [21] extended it for the binary features using Bernoulli distribution. However, no such formulation exists for the multinomial distribution. Moreover, such parametric link-based methods are never considered in the context of evolutionary clustering. We are motivated from both of these issues and propose a clustering method that exploits the links among the parameters of the multinomial distributions to analyze the temporal/evolutionary data.

Our overall contribution in this research is to propose a novel evolutionary clustering method based on multinomial mixture model. The highlights of our contributions include: (a) propose a formulation for a parametric link among multinomial distributions; (b) develop a novel evolutionary clustering method by exploiting the link parameters and (c) provide interpretation of the link parameters to describe cluster evolutions. First, we use synthetic data to evaluate and compare the proposed method w.r.t. the state-of-the-art methods. Next, we apply it to analyze the temporal dynamics of social media data obtained from the *ImagiWeb* project [36]. Results in Sec. 4 show that the proposed method is better than the state-of-the-art methods.

In the rest of the paper, we provide related background in Sec. 2, describe our proposed method in Sec. 3, present the experimental results and observations in Sec. 4 and finally draw conclusions in Sec. 5.

2. Background and related work

Evolutionary Clustering (ECL), also called *clustering over time*, aims to cluster the data that dynamically evolves over time [11]. ECL methods cluster the data by considering the temporal smoothness to reflect the long-term trends of the data while being robust to the short-term variations [11, 37, 12]. The demand and application of these methods are increasing rapidly in various domains. They have been successfully applied to analyze news [39], social media [23], stock price [37], photo-tag pairs [11], and documents [9].

Temporal/evolutionary data clustering has been addressed from several viewpoints in the literature, which naturally raises several task-specific notions about ECL. A distinction among them can be as follows: (1) clustering (2) monitoring and (3) interpreting. In the following paragraphs, we review relevant literature based on this distinction.

Following the definition of [11], the ECL method clusters data by considering the historic information and current data. Based on this definition, we do not consider the methods which do not take into account the historic information. Besides, in order to limit our focus on the parametric methods, we do not consider the methods from non-parametric Bayesian based approaches [38, 13, 22].

Numerous ECL methods have been proposed in the literature [11, 37, 12, 39, 23, 9]. [11] provided a generic framework and proposed different versions with the k-means and hierarchical clustering. It is based on optimizing a global cost function that consists of snapshot (static clustering) quality and history cost (temporal smoothness). [12] proposed two methods based on spectral clustering. In their approach, they added terms within the clustering cost functions in order to regularize the temporal smoothness. [37] recently proposed AFFECT, which performs adaptive evolutionary clustering by estimating an optimal smoothing parameter. It is extended with several static methods, such as k-means, hierarchical and spectral. A common property of these methods is that they are specialized for continuous data and hence may not be an appropriate choice for categorical data which is our concern in this research.

Dynamic Topic Model (DTM) is a well-known probabilistic method for analyzing temporal categorical data [9]. It extends the popular topic modeling method called Latent Dirichlet Allocation (LDA) [10]. It uses Dirichlet prior based smoothing, which sometime over-smooth the data. As a consequence, it may cluster the data samples with non co-occurring features in the same group [23]. This eventually causes DTM to underperform to cluster some classical non-textual temporal categorical data. Recently, [23] address this issue and proposed Temporal Multinomial Mixture (TMM). TMM extends the classical multinomial mixture model by incorporating temporal dependency into the relation between the data of current time epoch and

the clusters of the previous time epoch. MM is a standard probabilistic model, which has been widely used to cluster static discrete/categorical data [27, 34]. Similar to MM, TMM estimates model parameters using an Expectation Maximization (EM) algorithm. Although both DTM and TMM provide reasonable results to cluster temporal categorical data, they are unable to detect and provide any interpretation of the cluster evolutions, which is one of the main foci of this research. Indeed, TMM is more related to our proposed approach as we aim to establish parametric link among MMs at different time epochs.

The evolution monitoring task [35, 31, 14, 25] tracks clusters evolution by identifying the birth, death, split, merge and survival of clusters at different time. An external clustering method is first used at each time, e.g., [35] and [31] used the k-means method, whereas [25] used the neural clustering method. Afterward, the mapping among the clusters at different time is examined based on several heuristics. A different method, called label-based diachronic approach [25], exploits the MultiView Data Analysis among the cluster labels at different time. This approach constructs heuristics from features for monitoring cluster evolution. Our approach is different than the above methods, because: (a) we do not aim to propose a cluster monitoring method explicitly and (b) we do not use a static clustering method. Besides the above methods, [14] proposed a joint clustering-monitoring method which uses the cross association algorithm to cluster data and a bipartite graph to monitor evolution. For data clustering, they group the distinct features (word) in each cluster and hence features do not coexist in different clusters. This is different than us as we exploit all the features in order to provide a feature level interpretation for the evolution.

The task of evolution interpretation aims to explain the reason for the evolution of clusters at different time. It can be accomplished by explicitly analyzing the features. To this aim, [25] used the F-measures from individual features and construct a similarity report. In our work, this interpretation can be directly obtained from the link parameters which are estimated as a part of clustering. Therefore, unlike [25] , we do not need any external analysis of the features.

Based on the above distinctions from several viewpoints (clustering, monitoring and interpretation), we find that our method is more similar to the evolutionary clustering methods rather than the evolution monitoring methods. Therefore, we compare our method only with the relevant state-of-the-art evolutionary clustering methods, such as [37], [9] and [23].

Now we focus on the literature related to our proposal. The idea of parametric link in a transfer learning context [3] is inherited from the concept for Generalized Discriminant Analysis (GDA) [4]. GDA adapts the classification rule from a source population to a target population through a linear link map of their parameters. [4] proposed several models for GDA within the context of multivariate Gaussian distribution. Later, [21] extends the work of [4] for binary data using Bernoulli distribution [8]. We observe that these approaches can be exploited for developing an evolutionary clustering method by replacing the notion of source/target with different time epochs $t-1/t$. Besides, such development requires the derivation of the linear link for the multinomial distribution.

The multinomial distribution is a standard probability distribution for analyzing the discrete categorical data [1]. The Multinomial Mixture (MM) is a statistical model based on the multinomial distribution. It has been used for cluster analysis with discrete data [27, 1, 41, 34, 19]. [27] studied several Model-Based Clustering (MBC) methods with MM and experimentally compared them using different criteria such as clustering accuracy, computation time and number of selected clusters. [34] proposed a MBC method for MM which integrates both model estimation and selection task within a single EM algorithm. In their work, they extended the MBC strategy previously proposed by [15] and provided a formulation to compute the Minimum Message Length (MML) criterion for model selection. Most recently, [19] proposed a MBC method which performs simultaneous clustering and model selection using the MM. Their strategy performs similar task as [34] in a computationally efficient manner which has been previously proposed for the Gaussian distribution [17] and Fisher distribution [18]. Moreover, similar to [27], they provided a comparison among different model initialization and selection strategies. Following all of the above approaches [27, 34, 19], in this research we exploit the MBC framework to cluster discrete data with MM.

MBC [16, 28] is a well-established method for cluster analysis and unsupervised learning. It assumes a probabilistic model (e.g., mixture model) for the data, estimates the model parameters by optimizing an objective function (e.g., model likelihood) and produces probabilistic clustering. The Expectation Maximization (EM) [26] is mostly used in MBC to estimate the model parameters. EM consists of an Expectation step

(E-step) and a Maximization step (M-step) which are iteratively employed to maximize the log likelihood of the data. Initialization of the EM algorithm has significant impact on clustering results [26, 2]. The EM algorithm is sensitive to its initialization, because with different initializations it may converge to different values of likelihood function, some of which can be local maxima (i.e., sub-optimal results). In order to overcome this, numerous different initialization strategies are proposed and experimented in the relevant literature [6, 27, 2, 19]. Following recommendations, we use the small-EM [6, 7, 2, 19] method to initialize the MM parameters.

MBC has been commonly exploited to identify the best model for the data by fitting a set of models with different parameterizations and/or number of components and then applying a statistical model selection criterion [16, 5, 15, 28, 18]. In this paper, we apply this model fitting and selection strategy for two purposes: (a) to identify the parametric submodels (Section 3.3) and (b) to automatically select the number of components (Section 3.6).

3. Parametric Link Based Evolutionary Clustering

We adopt the parametric link approach [4, 21] for evolutionary clustering by assuming that the source samples are equivalent to the samples at time epoch t and target samples represent sample of time $t+1$. With this assumption, we incorporate linear link between multinomials at different time epoch. The algorithm for the proposed clustering method is presented in Algorithm 1.

3.1. Statistical model for evolutionary data samples

Let S^t be a set of samples corresponding to time t and S^{t+1} be a set from the next time $t+1$. We assume that while the cluster labels for S^t are known to us (estimated from $t-1$), labels of S^{t+1} are unknown.

Let S^t be composed of N^t pairs $(\mathbf{x}_1^t, \mathbf{z}_1^t), \ldots, (\mathbf{x}_{N^t}^t, \mathbf{z}_{N^t}^t)$ where $\mathbf{x}_i^t = \left\{x_{i,1}^t, \ldots, x_{i,D}^t\right\}$ is the D dimensional count vector of order V, i.e., $\sum_{d=1}^{D} x_{i,d}^t = V$ and $\mathbf{z}_i^t$ is the associated class label such that $\mathbf{z}_{i,k}^t = 1$ if the data belongs to cluster k with $k = 1, \ldots, K$ and $\mathbf{z}_{i,k}^t = 0$ otherwise. We assume that any sample $\mathbf{x}_i^t$ of S^t is an independent realization of the random variable $\mathbf{X}^t$ of distribution:

$$\mathbf{X}^t \sim \mathcal{M}(V, \boldsymbol{\mu}_k^t), \;\; k = 1, \ldots, K$$

with $\mathcal{M}(V, \boldsymbol{\mu}_k^t)$ is the V-order multinomial distribution with parameter $\boldsymbol{\mu}_k^t = (\mu_{k,1}^t, \ldots, \mu_{k,D}^t)$ which is formally defined[1] as [8]:

$$\mathcal{M}(\mathbf{x}_i | V, \boldsymbol{\mu}_k) = \binom{V}{x_{i,1}, x_{i,2}, \ldots, x_{i,D}} \prod_{d=1}^{D} \mu_{k,d}^{x_{i,d}} \tag{3.1}$$

here, $\boldsymbol{\mu}_k$ is the parameter of the multinomial distribution of k^{th} class with $0 \leq \mu_{k,d} \leq 1$ and $\sum_{d=1}^{D} \mu_{k,d} = 1$. Therefore, samples of the entire set S^t can be modeled with a mixture of K multinomials, also called Multinomial Mixture (MM) model, which has the following form[1]:

$$f\left(\mathbf{x}_i | \Theta_K\right) = \sum_{k=1}^{K} \pi_k \, \mathcal{M}(\mathbf{x}_i | V, \boldsymbol{\mu}_k) \tag{3.2}$$

In Eq. (3.2), $\Theta_K = \{(\pi_1, \boldsymbol{\mu}_1), \ldots, (\pi_K, \boldsymbol{\mu}_K)\}$ is the set of model parameters, π_k is the mixing proportion with $\sum_{k=1}^{K} \pi_k = 1$ and $\mathcal{M}(\mathbf{x}_i | V, \boldsymbol{\mu}_k)$ is the density function (Eq. (3.1)). Besides, we assume that the class label $\mathbf{z}_i^t$ is an independent realization of a random vector $\mathbf{Z}^t$, distributed according to 1-order multinomial:

$$\mathbf{Z}^t \sim \mathcal{M}(1, \boldsymbol{\pi}^t)$$

[1] In order to avoid redundancy, we do not use the superscript t for the notations in the generalized equations, such as (3.1, 3.2, 3.6, 3.7 and 3.8). Because, the definitions and derivations in these equations are time independent, i.e. remains same for any time instance t.

where $\boldsymbol{\pi}^t = \pi_1^t, \ldots, \pi_K^t$ is the mixing proportion of the model in Eq. (3.2).

The assumption of MM is similar for the samples of S^{t+1} with random variable $\mathbf{X}^{t+1}$ and parameter $\boldsymbol{\mu}_k^{t+1}$. However, for S^{t+1} the labels $\mathbf{z}_i^{t+1}$ of N^{t+1} pairs $(\mathbf{x}_1^{t+1}, \mathbf{z}_1^{t+1}), \ldots, (\mathbf{x}_{N^{t+1}}^{t+1}, \mathbf{z}_{N^{t+1}}^{t+1})$ are unknown. In the context of evolutionary clustering, our goal is to estimate the unknown labels $\mathbf{z}_i^{t+1}$ for $i = 1, \ldots, N^{t+1}$ using the information from S^t and S^{t+1} by establishing a link between $\boldsymbol{\mu}_k^t$ and $\boldsymbol{\mu}_k^{t+1}$.

3.2. Parametric link/relationship among temporal data

For random variables Y^t and Y^{t+1} distributed according to the Gaussian distribution, a linear distributional link exists (under weak assumptions) [4], which has the form: $Y^{t+1} \sim DY^t + b$, where D and b are the link parameters among the samples of different time epoch. For binary data the following distributional linear link among Bernoulli parameters (α^{t+1} and α^t with $0 \leq \alpha \leq 1$) is derived by [21]:

$$\alpha^{t+1} = \Phi\left(\delta\, \Phi^{-1}\left(\alpha^t\right) + \lambda\, \gamma\right) \tag{3.3}$$

where $\delta \in \mathbb{R}^+ \backslash \{0\}$, $\lambda \in \{-1, 1\}$ and $\gamma \in \mathbb{R}$ are the link parameters. Φ is the cumulative Gaussian function of mean 0 and variance 1, see Fig. 3.1. We can modify[2] the above formulation for multinomial parameters by considering two issues: (1) multinomial parameter $\boldsymbol{\mu}_k$ has similar property as $\boldsymbol{\alpha}_k$ except $\sum_{d=1}^{D} \mu_{k,d} = 1$ and (2) samples from X are not necessary to be binary, which makes λ as an unnecessary variable (it was introduced in [21] to handle binary observations). Considering these issues we can derive parametric link between $\boldsymbol{\mu}^t$ and $\boldsymbol{\mu}^{t+1}$ as:

$$\mu_{k,d}^{t+1} = \frac{\Phi\left(\delta_{k,d}\ \Phi^{-1}\left(\mu_{k,d}^t\right) + \gamma_{k,d}\right)}{\sum_{r=1}^{D} \Phi\left(\delta_{k,r}\ \Phi^{-1}\left(\mu_{k,r}^t\right) + \gamma_{k,r}\right)} \tag{3.4}$$

where $\delta_{k,d} \in \mathbb{R}^+ \backslash \{0\}$ and $\gamma_{k,d} \in \mathbb{R}$ are the link parameters. In Eq. (3.4), the combination of parameters $\delta_{k,d}$ and $\gamma_{k,d}$ for $\forall k, d$ is called a full model which is over-parameterized and may leads to ambiguity. Instead, we consider several sub-models with certain constraints on the parameters, see the following section.

3.3. Parametric sub-models

The idea of defining sub-models is frequent in Model-Based Clustering (MBC) [16]. We fit the evolutionary clustering model (Eq. (3.4)) with different sub-models and then select the best model using the Bayesian Information Criteria [33]:

$$BIC = -2L(\Theta) + \nu log\left(N^{t+1}\right) \tag{3.5}$$

where $L(\Theta)$ is the log-likelihood (Eq. (3.6)) value associated to the MM parameters of $t+1$, ν is the number of free parameters of the sub-model. These sub-models provide sufficient interpretation about the change in parameters from time t to $t+1$. Definition and interpretation of several basic sub-models, defined as pair $(\delta_{k,d}/\gamma_{k,d})$ are given below:

(M1) $1/0$**:** This model is constrained with $\delta_{k,d} = 1$ and $\gamma_{k,d} = 0$ for $\forall k, d$, i.e., $\nu = 0$. It indicates that the observations X^{t+1} can be modeled with $\mu_{k,d}^t$ and hence no evolution occurred.

(M2) $0/\gamma_{k,d}$**:** This model is constrained with $\delta_{k,d} = 0$ for $\forall k, d$, i.e., $\nu = K * D$. It indicates that the observations X^{t+1} should be modeled without considering $\mu_{k,d}^t$. This model should be selected when a new cluster evolved independently and does not consider any historical information. This is the most general model that can certainly fit the observations X^{t+1} to a MM most efficiently subject to a good initialization of the alternative iterative method. Several possible variations[3] of this model are: $0/\gamma$, $0/\boldsymbol{\gamma}_k$ and $0/\boldsymbol{\gamma}_d$.

(M3) $\delta_{k,d}/0$**:** This model is constrained with $\gamma_{k,d} = 0$ for $\forall k, d$, i.e., $\nu = K * D$. It indicates that $\mu_{k,d}^{t+1}$ are evolved through $\mu_{k,d}^t$ in a specific transformation space (inversed cumulative Gaussian). This model

[2] We cannot directly apply Eq. (3.3) for the multinomial distribution as it violates the constraint on parameter μ.

[3] Subscript k means cluster dependent and d means feature dependent. No subscription means a constant value for all clusters and features.

should be selected when true evolution occurred which can be explained in detail through certain belief on observed features and obtained clusters. Moreover, such a model can be plugged in with any other method in order to describe the cluster evolution. Several possible variations of this model are: $\delta/0$, $\boldsymbol{\delta}_k/0$ and $\boldsymbol{\delta}_d/0$. This model is equivalent to the fundamental unconstrained model assumed by [4].

(M4) $1/\gamma_{k,d}$**:** In this model, $\delta_{k,d} = 1$ for $\forall k, d$, i.e., $\nu = K * D$. This model does nearly similar task as model M3. It is relatively easier to fit through the additive term in the inverse cumulative Gaussian space. On the other hand, it is less expressive in terms of interpretation. Several possible variations of this model are: $1/\gamma$, $1/\boldsymbol{\gamma}_k$ and $1/\boldsymbol{\gamma}_d$.

3.4. Parameter estimation

In our proposed formulation of evolutionary clustering, we estimate two different types of parameters (see Eq. (3.4)): (1) MM model parameters: μ and π and (2) temporal link parameters: δ and γ. We estimate them in two steps. The first step consists of estimating μ and π (only for $t = 1$) for the observed samples of time t. In the second step, we estimate δ and γ. At any time epoch, we estimate the class labels $\mathbf{z}_i$ by *maximum a posteriori*.

3.4.1. Multinomial mixture parameters

We estimate the MM parameters using an Expectation Maximization (EM) algorithm that maximizes the log-likelihood value which has the following form[1]:

$$L\left(\Theta\right) = \sum_{i=1}^{N} log \sum_{j=1}^{K} \pi_j \mathcal{M}\left(\mathbf{x}_i|\boldsymbol{\mu}_j\right) \tag{3.6}$$

where N is the number of samples. In the Expectation step (E-step), we compute posterior probability as[1]:

$$\rho_{i,k} = p\left(z_{i,k} = 1|\mathbf{x}_i\right) = \frac{\pi_k \prod_{d=1}^{D} \mu_{k,d}^{x_{i,d}}}{\sum_{l=1}^{K} \pi_l \prod_{d=1}^{D} \mu_{l,d}^{x_{i,d}}} \tag{3.7}$$

In the Maximization step (M-step), we update π_k and $\mu_{k,d}$ as[1]:

$$\pi_k = \frac{1}{N} \sum_{i=1}^{N} \rho_{i,k} \quad \text{and} \quad \mu_{k,d} = \frac{\sum_{i=1}^{N} \rho_{i,k}\, \mathbf{x}_{i,d}}{\sum_{i=1}^{N} \sum_{r=1}^{D} \rho_{i,k}\, \mathbf{x}_{i,r}} \tag{3.8}$$

The E and M steps are iteratively employed until certain convergence criterion (difference of the log-likelihood values of successive iterations) is satisfied. The estimation of $\mu_{k,d}$ using Eq. (3.8) is only applicable for $t = 1$ due to the unavailability of any temporal information. For any time $t + 1$, when the link parameters are available, $\mu_{k,d}$ is estimated with Eq. (3.4).

3.4.2. Link parameters

Estimation of link parameters $\delta_{k,d}$ and $\gamma_{k,d}$ uses $\mu_{k,d}^{t}$ and the observed samples at time $t + 1$. Similar to [21], we use again an EM algorithm, but in which the M step is not explicit. Consequently, we employ an external optimization method such as an alternative iterative algorithm which consists of a succession, componentwise of the simplex method[4] [30]. In general, the starting point of the alternative algorithm corresponds to the case when $\mu_{k,d}^{t+1} = \mu_{k,d}^{t}$, i.e., $\delta_{k,d} = 1$ and $\gamma_{k,d} = 0$. However, in order to obtain a better estimate and save computation time [5], we apply an efficient approach, see Section 3.5.2.

[4]For the implementation, we used ***neldermead*** function of ***nloptr*** R package [40]. The lower and upper bounds were set to -2.5 and $+2.5$ respectively only for the $\gamma_{k,d}$ parameters.

[5]The simplex method requires a large number of iterations to converge.

Algorithm 1: Algorithm for clustering using Parametric Link among Multinomial Mixtures (PLMM).

Input: $\chi = \{S^t\}_{t=1,\dots,T}$, $S^t = \{\mathbf{x}_i^t\}_{i=1,\dots,N^t}$, $\mathbf{x}_i^t = \left\{x_{i,d}^t\right\}_{d=1,\dots,D}$, $x_{i,d}^t \in \mathbb{N}$

Output: Evolutionary clustering of χ with K classes and link parameters: $\delta_{k,d}^t$ and $\gamma_{k,d}^t$ $\forall k, d, t$.

```
foreach t do
    if t = 1 then
        Initialize π_k and μ_k for 1 ≤ k ≤ K using the small-EM procedure, see Section 3.5.1;
    end
    while not converged do
        {Perform the E-step of EM};
        foreach i and k do
            Compute ρ_ik = p(z_{i,k} = 1|x_i) using Eq. (3.7)
        end
        {Perform the M-step of EM};
        for k = 1 to K do
            if t = 1 then
                Update π_k and μ_k using Eq. (3.8)
            else
                Update π_k using Eq. (3.8)
                Compute δ_{k,d} and γ_{k,d}, see Sec. 3.4.2
                Update μ_k using Eq. (3.4)
            end
        end
    end
end
```

3.5. Parameters initialization

In the proposed clustering method (Algorithm 1), we need to initialize both the MM parameters $\Theta_K^{init} = \left\{(\pi_1^{init}, \boldsymbol{\mu}_1^{init}), \dots, (\pi_K^{init}, \boldsymbol{\mu}_K^{init})\right\}$ for time $t1$ and the link parameters (δ and γ).

3.5.1. Multinomial mixture parameters

Generally, the MM parameters are initialized randomly [27, 19]. However, with both synthetic and real data it has been demonstrated by [19] that, random initialization has its limitation w.r.t. the clustering performance and stability. Therefore, following [19], we initialize the model parameters using the small-EM procedure. This small-EM procedure consists of running multiple short runs of randomly initialized EM and then selecting the one with the maximum likelihood value. Here, short run means that the EM procedure does not need to wait until convergence and it can be stopped when a certain number of iterations is completed.

3.5.2. Link parameters

We propose an initialization procedure based on the predictive parameters set for next time epoch $\Theta_K^{pred} = \left\{(\pi_1^{pred}, \boldsymbol{\mu}_1^{pred}), \dots, (\pi_K^{pred}, \boldsymbol{\mu}_K^{pred})\right\}$. Let $\Theta_K^t = \{(\pi_1^t, \boldsymbol{\mu}_1^t), \dots, (\pi_K^t, \boldsymbol{\mu}_K^t)\}$ is the set of parameters for the current time (t) epoch. Our initialization procedure consists of the following steps:

- Step 1: estimate Θ_K^{pred} using data samples of next time X^{t+1} and an EM algorithm which is initialized with Θ_K^t.
- Step 2: compute $\delta_{k,d}^{init}$ and $\gamma_{k,d}^{init}$ for each k and d as:

$$\gamma_{k,d}^{init} = \Phi^{-1}\left(\mu_{k,d}^{pred}\right) \quad \text{for model M2} \tag{3.9}$$

$$\delta_{k,d}^{init} = \frac{\Phi^{-1}\left(\mu_{k,d}^{pred}\right)}{\Phi^{-1}\left(\mu_{k,d}^{t}\right)} \quad \text{for model M3} \tag{3.10}$$

$$\gamma_{k,d}^{init} = \Phi^{-1}\left(\mu_{k,d}^{pred}\right) - \Phi^{-1}\left(\mu_{k,d}^{t}\right) \quad \text{for model M4} \tag{3.11}$$

The Eq. (3.9), (3.10) and (3.11) are simply derived from Eq. (3.4) with the consideration that denominator is equal to 1, i.e., $\sum_{d=1}^{D} \mu_{k,d} = 1$ for $k = 1, \ldots, K$.

3.6. Varying number of clusters

The methodology presented in the previous sub-sections assumes the same number of clusters K for each time epoch. In this sub section, we propose an extension of it such that the method can handle varying K at different time, i.e., K^t and K^{t+1} may be different. To this aim, we modify the links initialization strategy (Section 3.5.2) in order to adapt the variability among $\Theta_{K^t}^{t}$ and $\Theta_{K^{t+1}}^{t+1}$. At time epoch t, this extended method requires additional information, such as: (a) number of clusters K^{t+1} and (b) cluster mapping between $\Theta_{K^t}^{t}$ and $\Theta_{K^{t+1}}^{t+1}$.

We adopted the method proposed by [19] with L-method [32] to select the number of cluster automatically at each time epoch. In order to initialize the link parameters, first we select the number of clusters K^{t+1} and obtain the predictive parameter set $\Theta_{K^{t+1}}^{pred}$. Next, for each cluster k in $\Theta_{K^{t+1}}^{pred}$ we find the corresponding cluster in $\Theta_{K^t}^{t}$ based on the minimum symmetric kullback leibler divergence (sKLD). sKLD among two clusters a and b is defined as [19]:

$$\begin{aligned} sKLD &= \frac{D_{KL}\left(\boldsymbol{\mu}_a, \boldsymbol{\mu}_b\right) + D_{KL}\left(\boldsymbol{\mu}_b, \boldsymbol{\mu}_a\right)}{2}, \quad \text{where} \\ D_{KL}\left(\boldsymbol{\mu}_a, \boldsymbol{\mu}_b\right) &= \sum_{d=1}^{D} \mu_{a,d}\, ln\left(\frac{\mu_{a,d}}{\mu_{b,d}}\right) \end{aligned} \tag{3.12}$$

After establishing the correspondences, we use Eq. (3.9), (3.10) and (3.11) to set the initial values of the link parameters. Finally, we estimate the link parameters following Section 3.4.2.

3.7. Interpretation of cluster evolution

The link parameters ($\delta_{k,d}$ and $\gamma_{k,d}$) along with the function Φ are the key to interpret the cluster evolution. Let us notice some basic interpretation of the values of these parameters for all feature d and cluster k:

- $\delta_{k,d} = 0$ means that $\mu_{k,d}$ (probability) at $t+1$ does not depend on t, whereas $\delta_{k,d} = 1$ (with $\gamma_{k,d} = 0$) means identical probability at two different times.
- $\delta_{k,d} \to 0$ and/or $\gamma_{k,d} \to \infty$ means that the distribution *tends to uniform* distribution.
- $\delta_{k,d} \to \infty$ and/or $\gamma_{k,d} \to -\infty$ means that the distribution tends to be *more concentrated* (Dirac distribution) at time $t+1$ in the feature which has the highest probability at time t.

In order to get further interpretation, we need to understand the multinomial parameters $\mu_{k,d}$ and the space spanned by the cumulative Gaussian Φ and its inverse Φ^{-1}. Let us consider an experiment of drawing V balls of $d = 1, \ldots, D$ different colors (represent features). After each draw, the color of the ball is recorded in a D dimensional count vector $\mathbf{x}_i$ and the ball is replaced. Therefore, at the end of i^{th} experiment $\mathbf{x}_{i,d}$ reveals the count of drawing the d^{th} colored ball. When a multinomial distribution is used to fit such experimental data, its parameter $\mu_{k,d}$ reveals the probability of drawing the d^{th} colored ball.

Now, let us consider Φ in Fig. 3.1 where the values along the Y-axis represent the possible values of $\mu_{k,d}^{t+1}$ (with $0 \leq \mu_{k,d}^{t+1} \leq 1$) and the X-axis represents the values of $\mu_{k,d}^{t}$ after transforming through Φ^{-1} function.

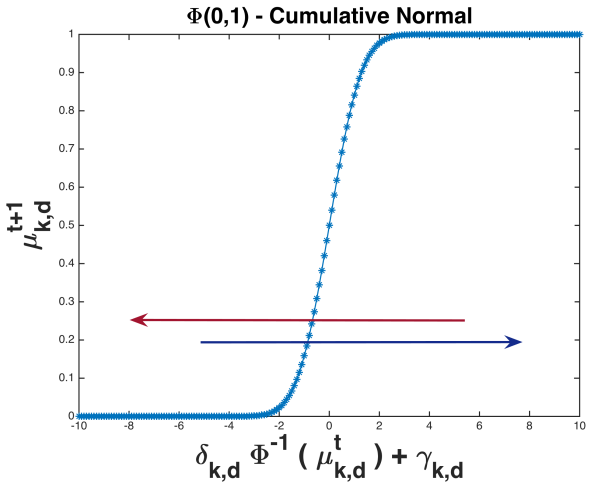

Figure 3.1: Illustrations of Cumulative Gaussian function and its relationship with the parameter change of multinomial distribution using Eq. (3.4). The arrows indicates the direction of changes in the inverse function space which eventually increase/decrease the probability.

Now, according to Eq. (3.4), cluster evolutions ($\mu^t_{k,d} \rightarrow \mu^{t+1}_{k,d}$) can be explained through multiplication (using $\delta_{k,d}$) and addition/subtraction (using $\gamma_{k,d}$) operations.

The values of $\gamma_{k,d}$ can certainly indicates the increase/decrease of the probability of certain feature (color) subject to the selection of sub-model **M4**. On the other hand if sub-model **M3** is selected, values of $\delta_{k,d}$ can explain the belief that $\mu^{t+1}_{k,d}$ should decrease if $\mu^t_{k,d} < 0.5$ and increase if $\mu^t_{k,d} > 0.5$. For example, let us consider that in a 2 colors (red and green) ball experiment the probability of the red color ball is changed from 0.8 (at time t1) to 0.7 (at time t2). Such a change can be explained with model **M3** with $\delta_{k,red} = 0.6$, which indicates that the belief is decreased at the next time. From the above discussions it is evident that the proposed method is capable to interpret the cluster evolutions up to the feature level.

4. Numerical experiments

We begin the experiments using simulated evolutionary data samples and evaluate w.r.t. the state-of-the-art methods. A characteristic comparison of different methods is presented in Table 1. For the simulated samples; we use the Adjusted Rand Index (ARI) [20] as a measure for evaluation. Next, we experiment and compare methods using real data. We use one of the real datasets experimented by [23]. We choose the *political opinion dataset* from the ImagiWeb project [36] as it consists of data from an interesting time period - during and after the election.

Table 1: Characteristic comparison of different state-of-the-art evolutionary clustering methods: Parametric Link among Multinomial Mixtures (PLMM, our proposed method), Temporal Multinomial Mixture (TMM) [23], Dynamic Topic Model (DTM) [9] and adaptive evolutionary clustering method (AFFECT) [37].

	PLMM	**DTM**	**TMM**	**AFFECT**
Data Type	Discrete	Discrete	Discrete	Continuous
Interpret Evolution	Yes	No	No	No

4.1. Simulated Data Samples

Following standard sampling methods we generate different sets $\{S^t\}_{t=1,\ldots,T}$ of simulated data for different time epochs. We draw a finite set of categorical samples (discrete count vectors) $S^t = \{\mathbf{x}_i\}_{i,\ldots,N^t}$ with different numbers (10, 20 and 40) of features (dimensions) D. These samples are issued from multinomial mixture (MM) models of $K = 3$ classes. We consider two different sets of samples:

- Samples with higher order of categorical count (***hos***) with $V \sim 1.5 * D$ with 3 time epochs each having different number of i.i.d. samples: $N^1 = 500$, $N^2 = 100$, and $N^3 = 200$. We also add noisy counts with these samples. These type of samples provides better resemblance with the MM parameters due to sufficient number of count in the observations. Practically, this is similar to the fact when the observations consists of data over longer period of time.

- Samples with lower order of categorical count (***los***) with $V \sim 0.7 * D$ with 5 time epochs each having different number of i.i.d. samples: $N^1 = 50$, $N^2 = 40$, $N^3 = 40$, $N^4 = 30$ and $N^5 = 20$. This type of samples are sparse and often difficult to distinguish among clusters. Practically, this is similar to the fact when the observations consists of data over shorter period of time.

The evolutionary data generation process consists of two steps: (1) determine MM parameters $\mu_{k,d}$ at each time epoch $t = 1, \ldots, T$ and (2) sample observations from the specified MM following assumption specified by [10]. For $t = 1$, we sample $\mu_{k,d}$ from a Dirichlet distribution and verify (separation w.r.t. the other clusters parameters [34]) it using the symmetric Kullback-Leibler Divergence value. For $t > 1$, we sample $\mu_{k,d}$ from $\mu_{k,d}^{t-1}$ using the MM link relationship defined in Eq. (3.4). This ensures that we maintain the temporal smoothness property [11, 37] of the evolutionary data samples. In order to use the link relationship, first we randomly select a model and then set the associated link parameters ($\delta_{k,d}$ and $\gamma_{k,d}$) within a pre-specified range of values.

To sample observations, first we choose the order V_k of each cluster. Our sampling procedure for each observation i at each time t follows the steps below:

- Choose a cluster $z_{i,k} = 1$ as: $\mathbf{z}_i \sim \mathcal{M}(1, \pi_1, \ldots, \pi_D)$, with, $\pi_d = \frac{1}{k}$.
- Choose the order τ_i of multinomial for $\mathbf{x}_i$ using Poisson distribution as: $\tau_i | z_{i,k} = 1 \sim \text{Poisson}(V_k)$.
- Draw sample $\mathbf{x}_i$ using multinomial distribution as: $\mathbf{x}_i | \tau_i, z_{i,k} = 1 \sim \mathcal{M}\big(\tau_i, \mu_{k,1}, \ldots, \mu_{k,D}\big)$.

Table 2: Simulated data evaluation and comparison using Adjusted Rand Index (ARI) [20]. Methods: PLMM (proposed), Dynamic Topic Model (DTM), Temporal Multinomial Mixture (TMM) and AFFECT with k-means. Datasets consist of different types (*hos* and *los*) of samples with different numbers (10, 20 and 40) of features. ***hos***: higher order samples and ***los***: lower order samples. **Boldfaced** indicate the best result and underlined numbers indicate second best. Values inside the parentheses provide the standard deviation of the ARI values.

	PLMM	**TMM**	**DTM**	**AFFECT**
10, *hos*	**0.91** (0.07)	0.86 (0.11)	0.79 (0.14)	0.43 (0.12)
10, *los*	0.81 (0.19)	**0.91** (0.1)	0.81 (0.1)	0.34 (0.11)
20, *hos*	**0.96** (0.05)	0.91 (0.1)	0.81 (0.18)	0.37 (0.11)
20, *los*	0.90 (0.18)	**0.98** (0.04)	0.95 (0.11)	0.35 (0.09)
40, *hos*	**0.97** (0.05)	0.92 (0.11)	0.48 (0.4)	0.33 (0.11)
40, *los*	0.93 (0.16)	**0.97** (0.05)	**0.97** (0.1)	0.36 (0.1)

We applied our proposed Parametric Link among Multinomial Mixtures (PLMM, Algorithm 1) clustering method on these simulated data using the basic sub-models defined in Sec. 3.3. Table 2 provides the results using the ARI [20] measure. Moreover, it provides a comparative evaluation w.r.t. other state-of-the-art methods (see comparison in Table 1): (a) Temporal Multinomial Mixture (TMM) [23] with smoothness parameter $\alpha = 1$; (b) Dynamic Topic Model (DTM) [9] with hyper-parameter $\alpha = 0.01$ and (c) Adaptive evolutionary clustering method (AFFECT[6]) [37] with k-means and Euclidean distance as a measure of similarity. We compute the average ARI of time $t = 2, \ldots, T$ (at $t = 1$ there is no evolution). Results in Table 2 w.r.t. ARI evaluation shows that:

[6]We experimented AFFECT with hierarchical and spectral clustering also. However, k-means provided the best results.

- PLMM (proposed) provides highest ARI for the ***hos*** samples and TMM [23] provides highest ARI for the ***los*** samples. These results are not surprising as both PLMM and TMM methods are specialized methods to cluster samples which are drawn from multinomial distributions.
- DTM [9] provides better results for ***los*** samples and higher dimensional data. This type of data is more likely to extract from text documents for which DTM was originally proposed.
- AFFECT [37] performs poorly compares to others for both types of sample. This is expected because of the similarity measure used in AFFECT is appropriate for continuous data.

Next, we test statistical hypothesis among PLMM, TMM and DTM using *two sample t-test* at the 5% significance level. The null hypothesis is that - the data in two results comes from independent random samples from normal distributions with equal means and equal but unknown variances. Results show that for all *hos* data the hypothesis is rejected with p-value<0.001. On the other hand, for the *los* data it is rejected only for 10 dimensional samples among the pairs (PLMM, TMM) and (DTM, TMM) with p-value<0.0001.

Next, we analyze the evolution of the clusters in terms of selected sub-models. Table 3 provides the rate of different selected models. We see that, for the ***hos*** data samples the model M4 $(1/\gamma_{k,d})$ is mostly selected. On the other hand, for the ***los*** data samples, different models M1: $(1/0)$, M4: $(1/\gamma_{k,d})$ and M3: $(\delta_{k,d}/0)$ are selected at certain rate. This observation confirms that PLMM successfully recovers the cluster evolutions with different models which were used to generate the simulated data. Interestingly, we observe that the model M2 $(1/\gamma_{k,d})$ is not selected which reflects the true fact that it was not considered to generate the simulated data samples. Now based on the selected model, we can provide further interpretation using $\delta_{k,d}$ and $\gamma_{k,d}$, see Sec. 3.3.

Table 3: Percentage of the selected models for the interpretation of evaluation. ***hos***: higher order (categorical count) samples and ***los***: lower order samples. **Boldfaced** indicate the highest rate.

	M1: $(1/0)$	**M4:** $(1/\gamma_{k,d})$	**M3:** $(\delta_{k,d}/0)$	**M2:** $(0/\gamma_{k,d})$
10, *hos*	0 %	**94 %**	6 %	0 %
10, *los*	15 %	38 %	**47 %**	0 %
20, *hos*	0 %	**92 %**	8 %	0 %
20, *los*	14 %	**43 %**	**43 %**	0 %
40, *hos*	0 %	**96 %**	4 %	0 %
40, *los*	4 %	37 %	**59 %**	0 %

Finally, we conduct experiments with varying number of clusters K at different time epoch. For this experiment, we use the same MM parameters which were used to generate the *hos* data samples. To ensure different K at different epoch, we randomly select a pair of time epochs and remove a cluster from one of them. Then, we generate $N^t = N^{t+1} = 1000$ synthetic data samples from them using the same procedure mentioned before. Applying the extension of PLMM method (Section 3.6) on these data provides the following results (ARI): 0.967 (0.09) for $d = 10$, 0.988 (0.04) for $d = 20$ and 0.986 (0.05) for $d = 40$. These results confirms that our proposed extension can cluster the synthetic data with varying K and provides reasonable accuracy.

4.2. Real data analysis: Opinion mining from twitter data

In order to challenge the applicability of the proposed method on real world data we focus on a relevant dataset which: (a) consists of discrete/categorical data and (b) can be divided into multiple meaningful timestamps. To this aim, we collected data from the *political opinion dataset* of the ImagiWeb[7] (IW-POD) project [36]. The motivation for choosing these data is that it consists of relatively lower number of features. Therefore, an evolution can be interpreted within a relatively easier and meaningful context.

[7] http://mediamining.univ-lyon2.fr/velcin/imagiweb/dataset.html

IW-POD consists of manually annotated tweets, from May 2012 to January 2013, related to two French politicians: Francois Hollande (FH) and Nicolas Sarkozy (NS). First, these tweets are annotated into 11 different aspects, such as Attribute (Att), Person (Per), Entity (Ent), Skills (Skl), Political line (Pol), Balance (Bal), Injunction (Inj), Projet (Pro), Ethic (Eth), Communication (Com) and No aspect detected (N/A). Afterward, each aspect is annotated with 6 opinion polarities, such as very negative (-2), negative (-1), no polarity (0), Null, positive (+1) and very positive (+2). For example, the tweet - *Sarko is more rational (orig: Sarko est plus rationnel)* is annotated with the aspect called *Person* and polarity +1. It is about NS and indicates that the user provides positive opinion with an emphasis on the personal attribute. Another example, the tweet - *Nicolas Sarkozy, the worst president of the Fifth Republic (Orig: Nicolas Sarkozy, le plus mauvais président de la Vème République)* is annotated with the aspect called *Skill* and polarity -1. It is a negative opinion about NS and indicates that the user emphasizes on the skill of NS.

In order to use these tweets for clustering, they are regrouped within the specified time epoch. Moreover, similar polarities are merged, e.g., two positives (+1 and +2) are merged into one as only positive (+). Therefore, each aspect consits of four polarities, such as positive (+), negative (-), zero (0) and undefined/null ($\emptyset$). As a consequence, finally each regrouped tweet represents the opinion of an user about a particular politician which is a 44 (11×4) dimensional vector of discrete data. In our experiment, we group opinions from IW-POD into three time[8] epochs: $t1$, $t2$ and $t3$, see Table 4 for details of the temporal data. Moreover, since the true number of clusters is unknown, we run clustering for different numbers of clusters ranging from 3 to 9.

Table 4: Details of the IW-POD dataset which is divided into three time periods. Each observation consists of a 44 dimensional discrete valued vector that encodes information about 11 different aspects each having 4 polarities.

Time stamp	**Time period**	**Significance**	**Num. opinions N. Sarkozy**	**Num. opinions F. Hollande**
t1	03/12 - 06/12	Before and After Election	1018	1168
t2	07/12 - 10/12	After Election	1067	1079
t3	11/12 - 01/13	After Election	1079	708

4.2.1. Comparison among different methods

We consider three different methods, Dynamic Topic Model (DTM) [9], Temporal Multinomial Mixture (TMM) [23] and Parametric Link among Multinomial Mixtures (PLMM), for a comparative evaluation of the performance on IW-POD dataset. These methods are selected based on their specialty to cluster discrete evolutionary/temporal data. We set 100 maximum number of iterations as the convergence criterion for all methods. Besides, we set the threshold log-likelihood difference values as 0.0001 for PLMM and TMM. The smoothness parameter α of TMM was set to 1. The DTM hyper-parameter α was set to 0.01. For the PLMM method, we consider the sub-models mentioned in Sec. 3.3.

IW-POD dataset does not provide ground truth cluster labels, due to which we were unable to evaluate clustering results with the known-labels based metric such as ***ARI***. In this context, we evaluate the methods using a well known likelihood related measure called *perplexity* on a held-out test set [29, 10]. ***Perplexity*** is a quantity originally used in the field of language modeling [29]. It measures how well a model has captured the underlying distribution of language. In clustering context, *perplexity* is defined as the reciprocal geometric mean of the per feature (word) log-likelihood of a test set, which is computed using the model parameters learned with a training set. Therefore, the *lower perplexity* value indicates that the estimated (trained) model performs *better* to fit the test data. ***Perplexity*** can be formally defined as [10]:

$$perplexity(X^{test}) = exp\left(-\frac{L\left(\Theta^{train}\right)}{\sum_{i=1}^{N^{test}} V_i}\right) \tag{4.1}$$

[8]The first round of the presidential election was held in 22/04/2012 and the second round run-off was held on 06/05/2012. Therefore, the data collected during this election period belong to time epoch $t1$.

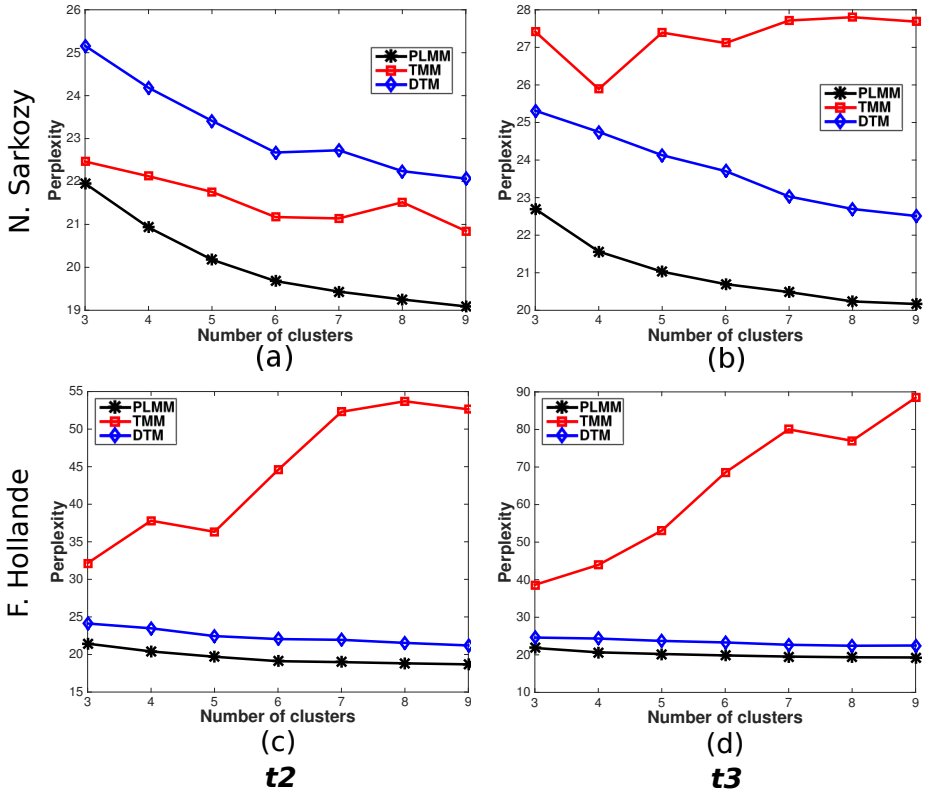

Figure 4.1: Comparison of different methods w.r.t. the *perplexity* values (***lower is better***) computed from the IW-POD data of two entities (row-1: Sarkozy and row-2: Hollande) and two time epochs (column-1: epoch $t2$ and column-2: epoch $t3$). Methods: Dynamic Topic Model (DTM) [9], Temporal Multinomial Mixture (TMM) [23] and our proposed Parametric Link among Multinomial Mixtures (PLMM) method.

where, V_i is the total number of feature counts (words for document) in observation i, $L\left(\Theta^{train}\right)$ denotes the log-likelihood of the test data set computed using the trained model parameters Θ^{train} and Eq. (3.6).

In our experiments, for each time epoch t, we compute *perplexity* from 5 folds of training-test data division and then take the average of 5 *perplexity* values as the final measure. For each fold, we used 80% data for training the model and obtain parameters Θ^{train} and the remaining 20% data to compute *perplexity* using Eq. (4.1). Fig. 4.1 illustrates the perplexity values computed from the data of two entities (row-1: Sarkozy and row-2: Hollande) and two time epochs (column-1: epoch $t2$ and column-2: epoch $t3$). Time epoch $t1$ is not considered because it does not reflect the link relationship and temporal aspect of data clustering.

Results in Fig. 4.1 show that, PLMM provides the best *perplexity* compared to DTM and TMM. This means that, compared to other methods, PLMM provides better fitting of the underlying multinomial distribution to the test data. The next best (3 out of 4) method is the DTM followed by the TMM. Indeed, the results from TMM are intuitive as the fitted models are highly influenced by the other cluster components (multinomial distributions) from the previous and next time epochs. In contrary, PLMM only consider the link from one cluster in the previous time epoch and fit the data accordingly.

Fig. 4.2 provides a visual illustration of clustering results obtained from the above three methods. This illustration is obtained by using the Multidimensional scaling [24] technique where the distance matrix among the observations is computed by first converting the count vectors into probabilities and then using the sKLD (Eq. 3.12) as a measure of distance. The clustering results are obtained with $K = 3$, time epoch $t2$ and the observations associated with the entity NS. From visual comparison among the plots in Fig. 4.2,

we can say that PLMM provides better separation than TMM and DTM. Indeed, this observation agrees with the numerical results obtained with the *perplexity* values in Fig. 4.1(a) for $K = 3$.

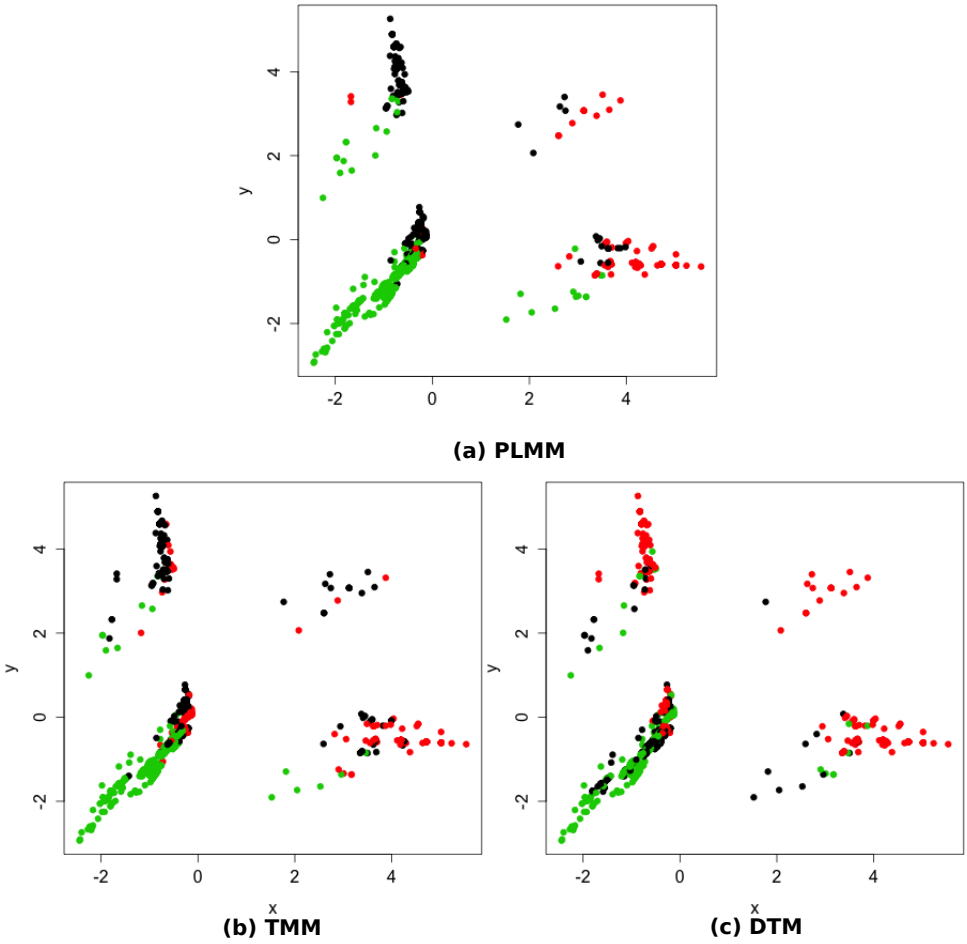

Figure 4.2: Illustration of clustering results visualized with Multidimensional scaling [24]. Methods: (a) proposed Parametric Link among Multinomial Mixtures (PLMM); (b) Temporal Multinomial Mixture (TMM) [23] and (c) Dynamic Topic Model (DTM) [9].

Next, we apply the extension of PLMM method (Section 3.6) with this dataset and observe the *perplexity* for time epochs $t2$ and $t3$. For the entity NS, we obtain average *perplexity* values as: $t2: 26.56$ and $t3: 25.06$ where average K^{t2} is 3 and average K^{t3} is 5. For the entity FH, we obtain average *perplexity* values as: $t2: 13.08$ and $t3: 5.17$ where average K^{t2} is 4 and average K^{t3} is 5. Compared to the results in Fig. 4.1 we see that, *perplexity* values increases (performance decreases) for entity NS and decreases (performance improves) for FH. Based on these observations, we can say that the extension of PLMM provides a good compromise in performance and works well for varying K at different epochs. We do not compare these results with the TMM and DTM methods as they work with fixed K for all time epochs.

Finally, let us focus on the interpretations of cluster evolutions in the IW-POD dataset. Table 5 provides the selection rate of different models at different time epochs (see Table 4 for details of time division). Listed rates provide us very interesting observations from which we can say that:

- The opinions about NS were evolving almost similar way during and after the election period. These evolutions can be interpreted through the belief on aspects using models M3:($\delta_{k,d}/0$) (93%) and

M4:$(1/\gamma_{k,d})$ (7%). This indicates that during $t1$-$t2$-$t3$ opinions about NS were changing slowly.

- Model M2:$(0/\gamma_{k,d})$ is selected for all clusters of opinions about FH during t1-t2. This means that the opinions change significantly between $t1$ and $t2$ period. From $t2$ to $t3$ (both after election period), opinions were evolving, which can be interpreted through the belief on the features with the models M4:$(1/\gamma_{k,d})$ (62%) and M3:$(\delta_{k,d}/0)$ (38%).

Table 5: Selection rate of different models (Sec. 3.3) for the IW-POD dataset at different time epochs (see Table 4 for details of time division).

	M1: $(1/0)$	**M4:** $(1/\gamma_{k,d})$	**M3:** $(\delta_{k,d}/0)$	**M2:** $(0/\gamma_{k,d})$
NS (t1-t2)	0 %	0 %	**100 %**	0 %
NS (t2-t3)	0 %	13 %	**87 %**	0 %
FH (t1-t2)	0 %	0 %	0 %	**100 %**
FH (t1-t2)	0 %	**62 %**	38 %	0 %

4.2.2. Cluster analysis, visualization and interpretation

In this section, we analyze the clustering results only from the PLMM method. In order to visualize the contents, we construct a histogram representation. It is constructed by counting the polarities (in vertical direction) w.r.t. each attribute (in horizontal direction). The color of the bars resembles the color of polarities. Fig. 4.3 and 4.4 illustrates examples of the clusters at different time epochs for the entities NS and FH respectively. These results are obtained by clustering data with $K = 3$. From both figures we observe that, at each time epoch the clusters have different histogram representations. Moreover, during different time epochs each cluster undergoes certain changes in different attributes and polarities. This demonstrates that PLMM method is able to provide sufficient inter-cluster variations (at each time) while respecting the temporal dynamics (during different time epochs).

An alternative and compact representation (w.r.t. the MM model parameters) of the clusters for NS is illustrated in Fig. 4.5(a) and 4.5(b). Similar to the examples of Fig. 4.3, this alternative representation demonstrate that, at a certain time epoch different cluster emphasizes on different aspects/polarities of an entity. Besides, the temporal changes of the clusters can be identified subsequently during different epochs by observing the increase/decrease of the probabilities. However, from the user's perspective, this representation may not be convenient to understand. Therefore, we use histograms for further analysis and use this compact representation for a different purpose.

Now, let us explain the semantics obtained from these clustering results. For brevity, here we denote a cluster as *cl.*. From Fig. 4.3 (clusters for NS) we see that, while *cl.* 1 and 3 emphasize on the negative (-) and positive (+) polarities respectively, *cl.* 2 emphasizes on a particular attribute. Naively we can say that, there are three groups of peoples: (a) the first group (*cl.* 1) provides negative opinions from various aspects, thus tends to hold a negative image about the entity; (b) the second group (*cl.* 2) particularly emphasizes on *Ethic* of the entity and mostly provide negative opinions and (c) the third group (*cl.* 3) can be seen as a contrary to the first group (*cl.* 1) as it tends to hold a positive image about the entity. Table 6 provides three examples of the tweets for time $t1$ and for each cluster about NS. We can realize that these tweets reflect the opinions which truly correspond to the groups obtained by the clustering method.

From temporal viewpoint, we observe several changes w.r.t. different aspects. In order to analyze the changes using histograms, we observe the height of histogram bar for each aspect. This height indicates the number of tweets/opinions corresponding to the related aspect. Let us consider an example of the aspect *Communication* which plays a certain role on clustering. We observe that: (a) for *cl.* 1, the total number of tweets related to the aspect *Communication* remains same during time $t1$ and $t2$ and reduces during $t2$ and $t3$; (b) for *cl.* 2, the total number of tweets related to this aspect reduces continuously and (c) for *cl.* 3, the total number of tweets related to this aspect reduces from $t1$ to $t2$ and remains same during $t2$ to $t3$. Moreover, a closer look on *cl.* 3 from $t2$ to $t3$ reveals an increase of positive opinions about the *communication* skill of the entity. Another example is the aspect called *Attribute*, whose height reduces

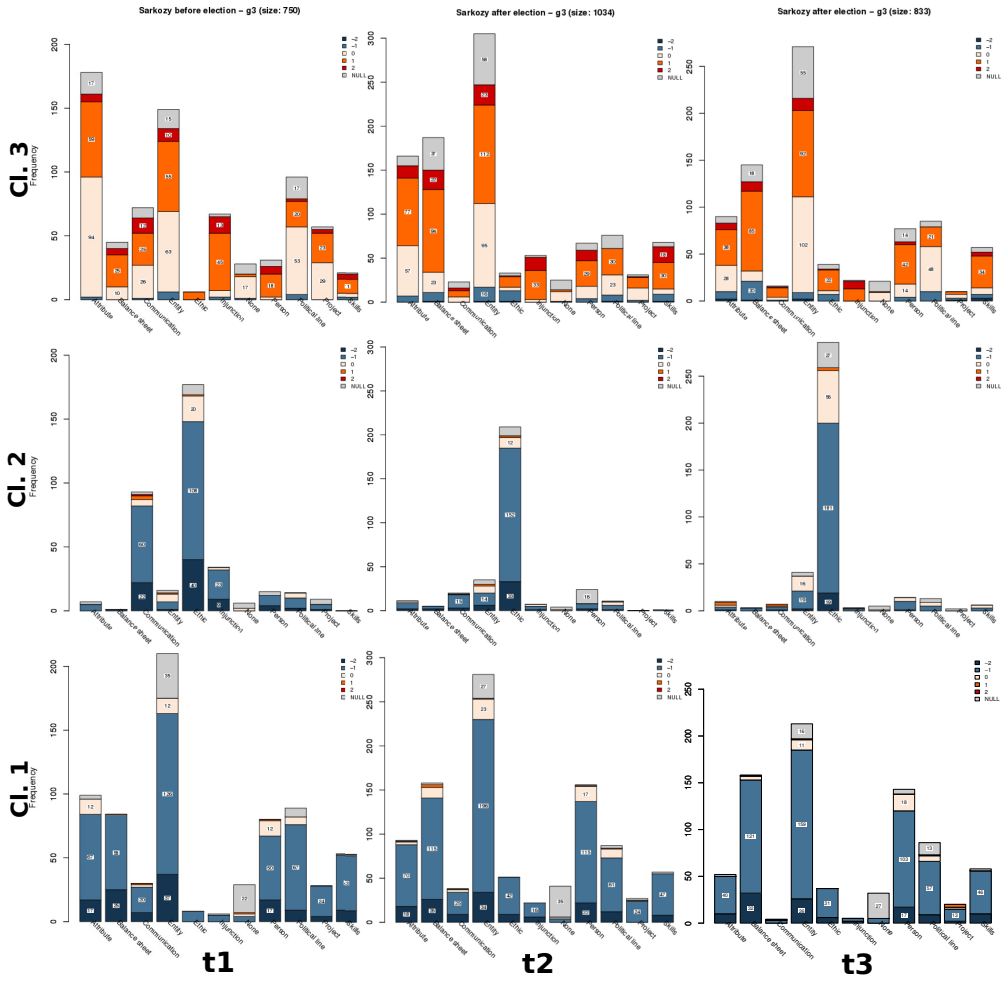

Figure 4.3: Illustration of the clustering results from PLMM methods for NS. Results obtained using $K = 3$ for three time epochs $t1$, $t2$ and $t3$. Each cluster is represented as a histogram constructed from the polarities of different aspects. The aspects are ordered from left to right as: (1) Attribute; (2) Balance sheet; (3) Communication; (4) Entity; (5) Ethic; (6) Injunction; (7) None; (8) Person; (9) Political line; (10) Project and (11) Skills. The polarities are colored and ordered from bottom to top as: -2 (dark blue), -1 (blue), 0 (light orange), 1 (orange), 2 (red) and NULL (grey). Each column represents clusters from a particular epoch. Each row represents a particular cluster in different epochs.

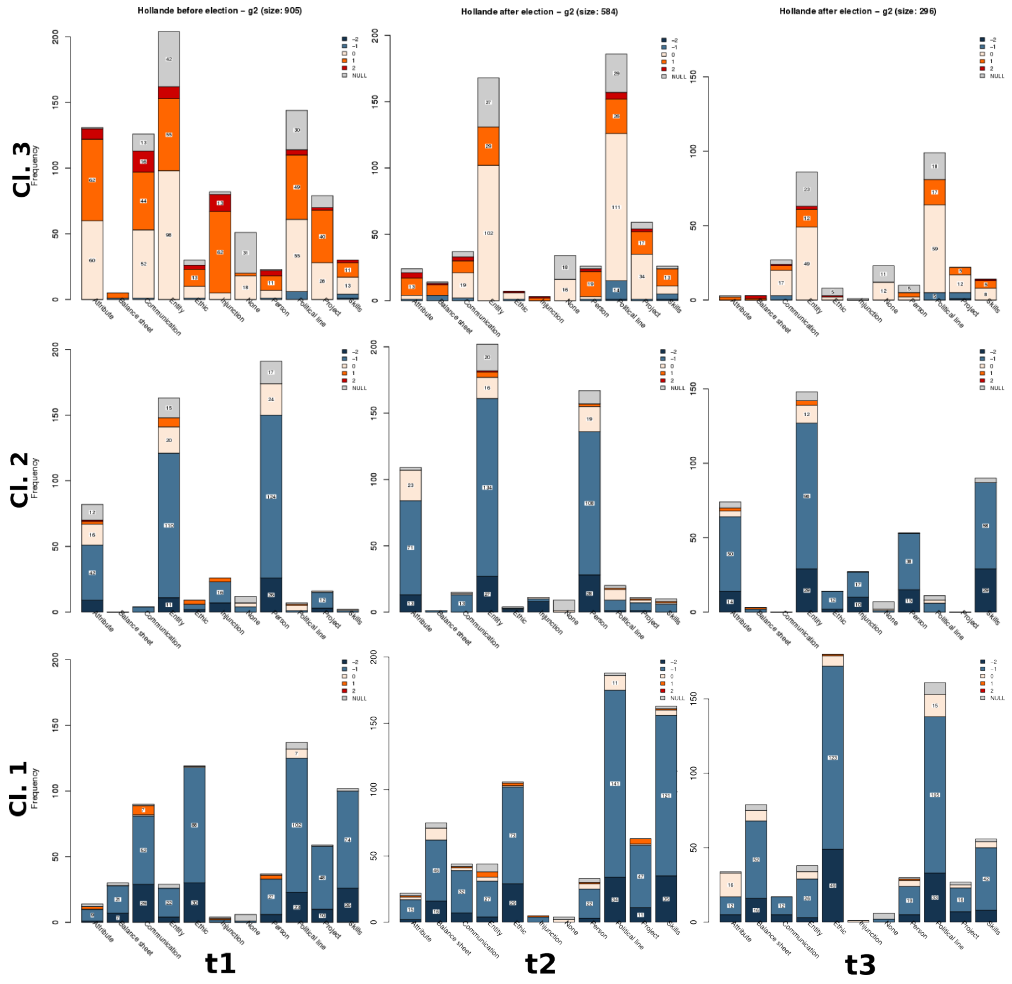

Figure 4.4: Illustration of the clustering results from PLMM methods for FH. Results obtained using $K = 3$ for three time epochs $t1$, $t2$ and $t3$. Each cluster is represented as a histogram constructed from the polarities of different aspects. The aspects are ordered from left to right as: (1) Attribute; (2) Balance sheet; (3) Communication; (4) Entity; (5) Ethic; (6) Injunction; (7) None; (8) Person; (9) Political line; (10) Project and (11) Skills. The polarities are colored and ordered from bottom to top as: -2 (dark blue), -1 (blue), 0 (light orange), 1 (orange), 2 (red) and NULL (grey). Each column represents clusters from a particular epoch. Each row represents a particular cluster in different epochs.

Table 6: Real twitter data examples of the 3 clusters at time $t1$ for entity NS. See Fig. 4.3 column 1 for the associated histograms.

	Cluster 1 (Generally Negative)
Ex. 1	**Orig:** Il veut desréférendums car... y a pas de pilote dans l'avion, dit-il: quel aveu! #Sarkozy#projet **Trans:** He wants referendumbecause... there is no pilot in the plane he says: what a confession! #Sarkozy#project
Ex. 2	**Orig:** Je ne voterais pas #Sarkozy ! " " Je ne voterais pas #Sarkozy ! **Trans:** I won't vote for #Sarkozy !" " I won't vote for #Sarkozy
Ex. 3	**Orig:**Nicolas Sarkozy, le plus mauvais président de la Vème République **Trans:** Nicolas Sarkozy, the worst president of the Fifth Republic
	Cluster 2 (Negative, specially "Ethic")
Ex. 1	**Orig:** Jamais un président n'a été cerné par tant d'affaires! demain ds @lematinch #Bettencourt #Sarkozy **Trans:** Never before a president was surrounded by so many cases!" tomorrow in @lematinch #Bettencourt #Sarkozy
Ex. 2	**Orig:** Une liste de condamnés de l'#UMP qui pourrait être bientôt complétée par les noms de #Sarkozy, #Copé, #Woerth **Trans:** A list of convicted people of #UMP soon completed by names such as #Sarkozy, #Copé, #Woerth (the "Bettencourt case" is a famous case in which Sarkozy was involved)
Ex. 3	**Orig:** Sarkozy-Kadhafi: la preuve du financement. Et l'urgence d'une enquête officielle #affairedetat **Trans:** Sarkozy-Kadhafi: the proof of funding. And the urge of an official enquiry #stateaffair (Kadhafi is another case in which Sarkozy was involved in some way)
	Cluster 3 (Generally Positive)
Ex. 1	**Orig:** N Sarkosy mots clé..challenge, défi, action, travail, réussite, formation, effort, individualisation ..France Forte. Europe Forte #NS2012 **Trans:** N Sarkozy keywords..challenge, défi, action, work, success, training, effort, individualization ..Strong France. Strong Europe #NS2012
Ex. 2	**Orig:** merci N.Sarkozy pour tout tu restera pour toujour mon Hero merci. merci **Trans:** Thank you N.Sarkozy for all you will stay my hero forever thanks. thanks
Ex. 3	**Orig:** Sarko est plus rationnel.. **Trans:** Sarko is more rational..

continuously with time for both *cl.* 1 and 3. Similarly, from an analysis of the height of histogram bars in Fig. 4.4 (clusters for FH) we see that, the aspects called *Entity*, *Ethic*, *Political line*, *Skills* and *Communication* play certain role to describe the image of FH. For example, the tweet - *Holland would remove the word "race" in the Constitution (orig: Hollande supprimerait le mot "race" dans la Constitution)* from time $t1$ and *cl. 3* is annotated with the aspect called *political line* and polarity *+1*. Another tweet - *Holland and Netanyahu evoke the struggle against anti-Semitism (orig: Hollande et Netanyahou évoquent la lutte contre l'antisémitisme)* has the same annotation which is from the same cluster but from time $t3$. These two examples reveal the importance of the aspect *political line* for keeping the similar opinions into the same group at different time. The above observations clearly indicate that, for different groups of people different aspects has certain importance at different time. Therefore, an analyst can retrieve the most prominent aspects from people's opinion about an entity at a particular time or within a certain range of time periods.

Besides the above interpretation of the clustering results, an analyst can obtain more information from the PLMM clustering results via the link parameters ($\delta_{k,d}$ or $\gamma_{k,d}$). After analyzing the links among MM parameters we notice that they are able to provide a compact explanation about the temporal changes during

two time epochs. Fig. 4.5 illustrates an example for entity *NS* from time $t1$ to $t2$ with 3 clusters, see column 1 and 2 of Fig. 4.3 for corresponding histograms. Fig. 4.5(a) and Fig. 4.5(b) illustrates the MM parameters (probability of aspect-polarity features) and Fig. 4.5(c) provides a compact representation about the cluster evolutions using the values of $\delta_{k,d}$. To better understand this representation in Fig. 4.5(c), we transform the link values as 0 (no change), -1 ($\delta_{k,d} < 0.9$, belief increases) and +1 ($\delta_{k,d} > 1.1$, belief decreases). In the context of the examples from the IW-POD, we can explain belief as: probability of a feature at time $t+1$ is increased from its probability at time t. Therefore, the belief indicates the relative significance of a particular feature w.r.t. time. An increase in the belief means that users tend to be more attracted by it. Following this, if a feature probability is nearly same at two different times then belief remains unchanged. In Fig. 4.5, we highlight the effect of a particular aspect, called *Communication* (*Com*), and observe its contribution for cluster evolution. From Fig. 4.5 (a) and (b) we see that, from time $t1$ to $t2$ the probabilities are decreased mostly for *cl.* 2 and 3. This means that, either the users from these clusters loose interest to discuss about *Com* and focus on other aspects, or those users disappeared at time $t2$. Similar to *Com*, we can observe other aspects such as *Eth* (*cl.* 1 and *cl.* 3) and *Ent* (*cl.* 2 and *cl.* 3) which causes cluster evolution in this example of Fig. 4.5.

Let us analyze examples from real twitter data and observe them w.r.t. Fig. 4.5. If we look at *cl.* 3 at $t1$ (before election), the most likely features are often positive and it is clear that it gathers people in favor of NS. The prominent aspects are *Att* (positive and neutral), *Ent* (positive) and *Inj* (positive), such as in the tweet - *40 people @youngpop44 will be present at the great gathering in Place #Concorde for supporting @NicolasSarkozy ! #StrongFrance #NS2012"*. This cluster slightly changes later at $t2$ (just after election) towards *Att* (positive), *Ent* (positive) and *Bal* (positive). The shift from *Inj* to *Bal* is clearly visible on Fig. 4.5(c), third row: black color for *Inj* means a decrease of attention whereas white color for *Bal* means there are relatively more comments on the balance sheet of NS. Hence, the following message shows some nostalgia felt by many militants: *Whatever the opinion of FH, NS has been a great president. FH can deconstruct all the reforms, we will never forget!*. To sum up, the δ parameter helps us to focus on what are the main changes, even though the observation could have been drawn among the other aspects. Following the same reasoning, all polarities targeting the aspect *Com* are black, which proves that the performances of the politician in the media (e.g., TV, newspapers) are less important once the election is over.

Observations from numerous experiments reveal that, besides performing evolutionary clustering on the temporal data, PLMM also provide reasonable interpretation for the evolutions, thanks to the link parameters. Indeed, this clearly distinguishes PLMM from the rest of the state-of-the-art methods. Moreover, we notice that the interpretability of PLMM (using Eq. 3.9, 3.10 and 3.11) can be separated out and externally plugged in with the results from any other discrete data clustering methods.

5. Conclusion and Future Perspectives

Over the years, a large number of temporal data analysis methods have been proposed in several domains. In this paper, we only focused on the particular clustering methods which have been used for discrete data clustering and which are based on the assumption of the multinomial distribution.

We proposed an unsupervised method (i.e., no training from labeled data) for analyzing the temporal data. The core element of our proposal is the formulation of parametric links among the multinomial distributions. Computations of these links naturally cluster the evolutionary/temporal data. Furthermore, these links can provide interpretation for cluster evolution and also detect clusters evolution in certain cases. For experimental validation, we extensively used synthetic dataset and evaluated using the *Adjusted Rand Index*. As a practical application, we applied it on a dataset of political opinions and evaluated using *Perplexity* measure. Results show that the proposed method, called PLMM, is better than the state-of-the-art. Moreover, it provides an additional advantage through the link parameters in order to interpret the changes in clusters at different time. We also provide an extension of the proposed method for dealing with varying number of clusters which is not addressed by most of the recent methods.

Monitoring/tracking cluster evolution is an interesting issue which we do not explicitly and extensively manage in our proposed method, because it is not a primary objective in this paper. Yet, we can partially achieve this task by using certain information (parametric sub-models, see 3.3) which are naturally integrated

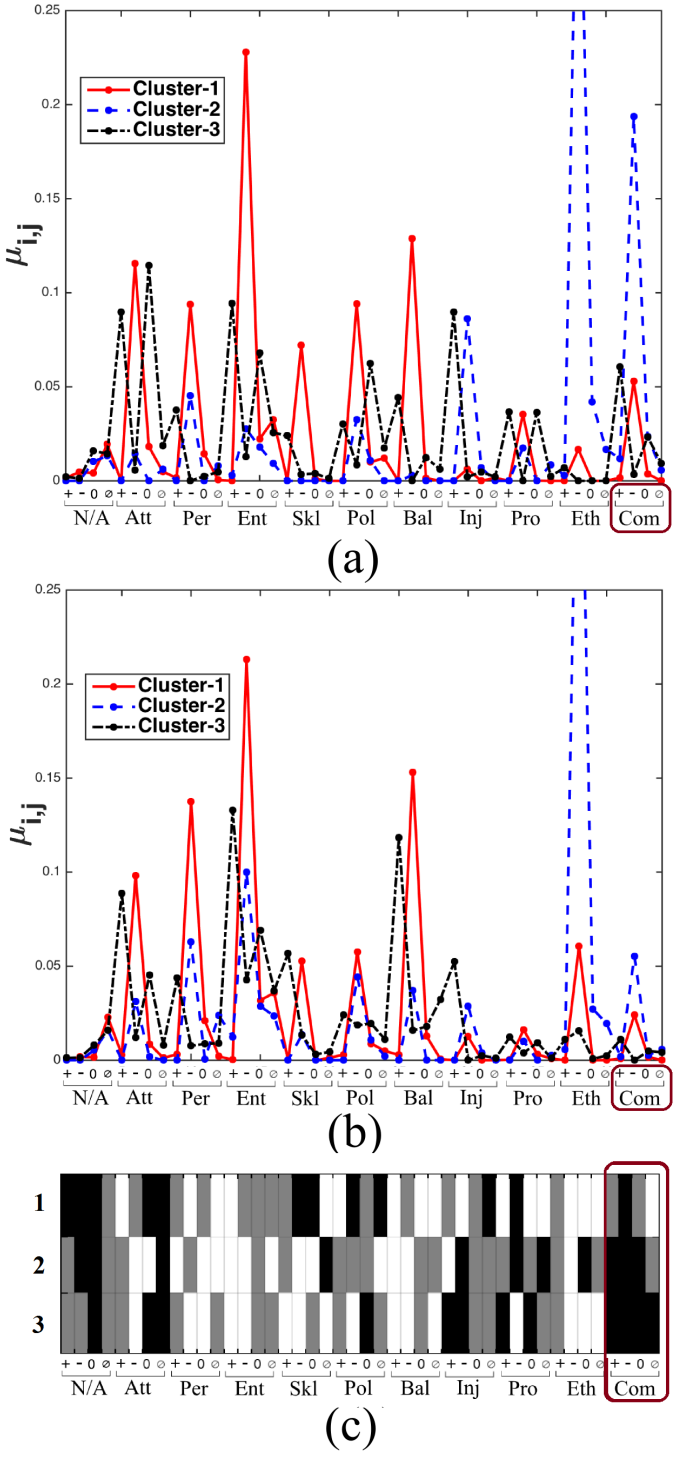

Figure 4.5: Example of evolution interpretation using link parameter $\delta_{k,d}$ for *NS during t1 to t2 with 3 clusters*. (a) MM parameters $\mu^{t1}_{k,j}$ at time *t1* (b) MM parameters $\mu^{t2}_{k,j}$ at time *t2* (c) Link parameters $\delta_{k,j}$ between time *t1* and *t2*. In (c), for each cluster (row-wise), brighter/white color indicates the prior belief about features (aspect-polarity) increases, darker/black color indicates the prior belief about features decreases and grey color indicates the prior belief about features remains same.

with our proposed method. That means, our proposed method can be used only as a detector of cluster evolution. At present, we consider the complete monitoring task as a future work. We believe that, an extension of several existing work can be added with our method to completely deal with this issue. For example, we can exploit[9] MEC [31] which is a cluster evolution monitoring method for continuous data. Besides, we can use *label-based diachronic approach* [25] by externally providing our clustering results as an input to it.

Computational complexity is a concern for the proposed method and can be considered as a limitation. From a decomposition of the computational time, we observe that most of the time is consumed by the optimization procedure (*neldermead* simplex method). In future, a better optimization method can be incorporated to address this issue. Moreover, the time can be further reduced by eliminating the parametric sub-models which are experimentally found as redundant.

Although we demonstrated the effectiveness of the proposed method only for political opinion dataset, we believe that it will be equally effective for different datasets that consist of the form of categorical data.

6. Acknowledgements

This work is funded by the project ImagiWeb ANR-2012- CORD-002-01.

References

[1] Agresti, A., 2002. Categorical data analysis. 2nd ed., John Wiley & Sons.
[2] Baudry, J.P., Celeux, G., 2015. Em for mixtures-initialization requires special care. Statistics and Computing 25, 713–726. doi:10.1007/s11222-015-9561-x.
[3] Beninel, F., Biernacki, C., Bouveyron, C., Jacques, J., Lourme, A., 2012. Parametric link models for knowledge transfer in statistical learning. Knowledge Transfer: Practices, Types and Challenges. Nova Science Publishers.
[4] Biernacki, C., Beninel, F., Bretagnolle, V., 2002. A generalized discriminant rule when training population and test population differ on their descriptive parameters. Biometrics 58, 387–397.
[5] Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE TPAMI 22, 719–725.
[6] Biernacki, C., Celeux, G., Govaert, G., 2003. Choosing starting values for the em algorithm for getting the highest likelihood in multivariate gaussian mixture models. Computational Statistics & Data Analysis 41, 561–575.
[7] Biernacki, C., Celeux, G., Govaert, G., Langrognet, F., 2006. Model-based cluster and discriminant analysis with the mixmod software. Computational Statistics & Data Analysis 51, 587–600.
[8] Bishop, C.M., et al., 2006. Pattern recognition and machine learning. volume 4. springer New York.
[9] Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models, in: Proc. of the Int Conf on Machine Learning, ACM. pp. 113–120.
[10] Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022.
[11] Chakrabarti, D., Kumar, R., Tomkins, A., 2006. Evolutionary clustering, in: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM. pp. 554–560.
[12] Chi, Y., Song, X., Zhou, D., Hino, K., Tseng, B.L., 2009. On evolutionary spectral clustering. ACM Trans. on Knowledge Discovery from Data 3, 17.
[13] Dubey, A., Hefny, A., Williamson, S., Xing, E.P., 2013. A nonparametric mixture model for topic modeling over time., in: SDM, SIAM. pp. 530–538.
[14] Ferlez, J., Faloutsos, C., Leskovec, J., Mladenic, D., Grobelnik, M., 2008. Monitoring network evolution using MDL, in: IEEE Int. Conf. on Data Engineering, IEEE. pp. 1328–1330.
[15] Figueiredo, M.A.T., Jain, A.K., 2002. Unsupervised learning of finite mixture models. IEEE TPAMI 24, 381–396.
[16] Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97, 611–631.
[17] Garcia, V., Nielsen, F., 2010. Simplification and hierarchical representations of mixtures of exponential families. Signal Processing 90, 3197–3212.
[18] Hasnat, M., Alata, O., Trémeau, A., 2015a. Model-based hierarchical clustering with bregman divergences and fishers mixture model: application to depth image analysis. Statistics and Computing , 1–20doi:10.1007/s11222-015-9576-3.
[19] Hasnat, M.A., Velcin, J., Bonnevay, S., Jacques, J., 2015b. Simultaneous clustering and model selection for multinomial distribution: A comparative study, in: Advances in Intelligent Data Analysis XIV. Springer.
[20] Hubert, L., Arabie, P., 1985. Comparing partitions. Journal of classification 2, 193–218.
[21] Jacques, J., Biernacki, C., 2010. Extension of model-based classification for binary data when training and test populations differ. Journal of Applied Statistics 37, 749–766.

[9] We conducted some initial experiments and found that this approach is applicable up to certain extent and should be further improved to use in our case, e.g., extend it with appropriate distance computation (e.g., using sKLD).

[22] Kharratzadeh, M., Renard, B., Coates, M., 2015. Bayesian topic model approaches to online and time-dependent clustering. Digital Signal Processing .
[23] Kim, Y.M., Velcin, J., Bonnevay, S., Rizoiu, M.A., 2015. Temporal multinomial mixture for instance-oriented evolutionary clustering, in: Proc. of the European Conference on Information Retrieval, pp. 593–604.
[24] Kruskal, J.B., Wish, M., 1978. Multidimensional scaling. volume 11. Sage.
[25] Lamirel, J.C., 2012. A new approach for automatizing the analysis of research topics dynamics: application to optoelectronics research. Scientometrics 93, 151–166.
[26] McLachlan, G.J., Krishnan, T., 2008. The EM algorithm and extensions. Wiley series in probability and statistics. 2. ed ed., Wiley.
[27] Meilă, M., Heckerman, D., 2001. An experimental comparison of model-based clustering methods. Machine Learning 42, 9–29.
[28] Melnykov, V., Maitra, R., 2010. Finite mixture models and model-based clustering. Statistics Surveys 4, 80–116.
[29] Murphy, K.P., 2012. Machine learning: a probabilistic perspective. The MIT Press.
[30] Nelder, J.A., Mead, R., 1965. A simplex method for function minimization. The computer journal 7, 308–313.
[31] Oliveira, M.D., Gama, J., 2010. MEC-monitoring clusters' transitions., in: STAIRS, pp. 212–224.
[32] Salvador, S., Chan, P., 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms, in: IEEE Conf. on Tools with Artificial Intelligence, pp. 576–584.
[33] Schwarz, G., et al., 1978. Estimating the dimension of a model. The Annals of Statistics 6, 461–464.
[34] Silvestre, C., Cardoso, M.G., Figueiredo, M.A., 2014. Identifying the number of clusters in discrete mixture models. arXiv preprint arXiv:1409.7419 .
[35] Spiliopoulou, M., Ntoutsi, I., Theodoridis, Y., Schult, R., 2006. MONIC: modeling and monitoring cluster transitions, in: Proc. of the ACM SIGKDD Int conf. on Knowledge discovery and data mining, ACM. pp. 706–711.
[36] Velcin, J., Kim, Y., Brun, C., Dormagen, J., SanJuan, E., Khouas, L., Peradotto, A., Bonnevay, S., Roux, C., Boyadjian, J., et al., 2014. Investigating the image of entities in social media: Dataset design and first results, in: Proc. of Language Resources and Evaluation Conference (LREC).
[37] Xu, K.S., Kliger, M., Hero Iii, A.O., 2014. Adaptive evolutionary clustering. Data Mining and Knowledge Discovery 28, 304–336.
[38] Xu, T., Zhang, Z., Yu, P.S., Long, B., 2008. Dirichlet process based evolutionary clustering, in: Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on, IEEE. pp. 648–657.
[39] Xu, T., Zhang, Z., Yu, P.S., Long, B., 2012. Generative models for evolutionary clustering. ACM Trans. on Knowledge Discovery from Data 6, 7.
[40] Ypma, J., 2014. Introduction to nloptr: an r interface to nlopt .
[41] Zhong, S., Ghosh, J., 2005. Generative model-based document clustering: a comparative study. Knowledge and Information Systems 8, 374–384.