Bayesian Model-Based Clustering for Community Detection

Alessandro Mirone - 966880

22/4/2022

UNIVERSITÀ DEGLI STUDI DI MILANO

Abstract

This paper focuses on the implementation of Bayesian analysis for grouping individuals based on what of their socio-political stance is expressed on social media. Through the use of Bayesian model-based clustering, it's possible to infer information about the connection between people's own beliefs and their membership to communities that share their same stance regarding social and political issues, in particular the probability of being in one such community given one's digital behaviour. It's also possible to draw conclusions about the characteristics of the groups, like their cohesion or the relative distance from other groups in terms of opinions. Most importantly, this work provides an efficient methodology using techniques from Bayesian Statistics and Computer Science for a Sociological analysis traditionally carried out with qualitative tools, connecting in the process sociological theory on digital practices and empirical evidence coming from a quantitative analysis of real data.

Introduction

Modern Sociology theory highlights the importance of digital practices in the construction of the Self and of the relations that one exerts throughout its life towards other people. (*Couldry*,2012; *Castells*,1996 & 2009). Since the works of social scientists the like of P. Bourdieu, it is believed that practices are defined by what we could refer to as "Habit" (*Habitus, Bourdieu*, 1979), a sort of social etiquette that is assimilated through relations with others and inherent to social groups; in fact, sharing the same Habit is both a prerequisite for people to recognize themselves as part of a certain social group and a byproduct of the culture -or subculture- which that same group heralds. (*Tajfel*,1978; *Goffman*,1959; *Mead*, 1934) Social communities themselves are defined in relation to the culture in which they develop; be that an opposing relation or a positive one, either way they understand themselves thanks to an exchange with other cultures -particularly the dominant ones, relatively to the space and time in which the groups are born- as do the people that compose those communities (*Schütz*, 1944). It exists a circular motion between practices and Habit: as an ensemble of practices define a Habit, so a certain social group identified with its Habit reproduce its cultural identity through the implementation of certain practices. Given that Habits are culturally defined objects, also practices are born from the same origin. Moreover, because of the connection between Habits/practices and the cultural identity of the groups, it's possible to make inference about the membership of an individual by observing its practices (as Habits are not directly observable). Of course, practices are not exclusive to some communities; many are shared between almost all social groups, so the choice of the objects to analyze is of the utmost importance. In this regard, qualitative analysis has found that the most distinctive practices are those that regard the consumption of cultural products (*Bourdieu,1979*). At this stage it's necessary to dwell on the definition of "Practice" and on what their cultural origin imply: we can define a practice as a set of behaviours, actions and beliefs that are expressed by an individual or a collective for a specific purpose; an example that may shed some light on the role of practices in defining the social framework of a community is the act of praying. This action originates from a specific cultural framework and replicates it through its implementation by the individuals that take part in it. However, the same individuals would perform that action only when they feel the need to do so, and won't pray while at work, for example, or while they go shopping, as those social frameworks are not coherent (relatively with the culture of their religious group) with the act; those situations require other practices, notably the ones that are suggested by the Habit of the individual performing them. In this sense, consumption of cultural products is really emblematic, as people with different Habits will almost certainly enjoy different products (*Bourdieu*,1979). It's also useful to clarify that as singular practices are not exclusive to certain Habits, individuals also are not members of just one group throughout their life; it's the set of practices and the decision on how and when to carry them out that is exclusive to the Habit, and the Habit identifies a social group. This decision is driven by the individuals' understanding of the social framework in which they find themselves at the moment of the choice (*Goffman*,1959). Since the ability to understand such frameworks is also culturally learned -exactly through practices-, it is typical to the different social groups. Following H. Garfinkel definition, we call "Ethnomethods" the way in which individuals understand the social frameworks and interpret the reality of everyday life situations (*Garfinkel*, 1967). These objects complete the picture, as it can now be seen how interpretations of reality -thus the realization of the Self- are culturally learned objects specific to social groups, and thus informed by a set of practices and a Habit; most importantly for the scope of this paper, we can see how opinions -that are, in fact, interpretations of a relative reality- are shaped by practices and Habits. Given this epistemic framework, it's possible to regard practices as indicators of social groups membership and the expression of opinions as a practice in itself informed by the Habit of a certain social group. In particular, this work studies digital practices (*Couldry*, 2012), focusing on one: the act of commenting on social platforms -In this case, Twitter-. To do so and inspect the probability of being in a certain group having observed the expression of a culturally learned opinion by an individual, the method chosen was that of model based clustering; this technique is well established as a tool for social sorting (*Lyon*, 2003) for marketing and security purposes. This paper does conceptually the same, but applying a Bayesian approach for model based clustering based on the recent works in the field (*Fraley & Raftery*, 2005; *Lau & Green*, 2007) and to a different scope: implementing a statistical quantitative method and a statistical learning algorithm for investigating the findings from sociological theory. In particular, this work follows the 2005 paper from Fraley and Raftery for estimating the model, implements the Expectation Maximization algorithm (EM) to find the best model's parameters given the assumption that data are generated from a finite mixture of multivariate Gaussian distributions and assigns each observation to a component of said mixture. This allows to make inference about the probability of being in one group or another given certain coordinates as well as providing general information about the groups' mean vectors and variance-covariance matrices' shapes, volumes, and orientations.

Data description

The first step in the analysis is the choice of the indicators of individuals' membership to unobserved social communities. Given the philosophical framework discussed above, it was chosen to evaluate one particular digital practice, that of *commenting* (or *Posting*)(*Couldry*, 2012); the statistical analysis was carried out over 15927 users of the social media platform Twitter, which Tweets (digital comments) were downloaded from the social network between December 2021 and January 2022. The choice of the digital platform gave the opportunity to focus on certain objects: *Hashtags.* These brief sentences -or just one word singletons- are meant to express the users' stance over a particular social fact and so configure themselves as the optimal indicator for community membership, as they can both be seen as cultural products which use is incentivized by the Habit of the community, and as digital practices (again, that of *commenting* plus the one of *sharing*) (*Couldry*, 2012). To build an indicator capable of localizing individuals among the social space, first it's necessary to define some dimensions, which will be later coded into variables. In particular:

- Racial: a dimension representing individuals' believes on racial equality
- Activism: a dimension representing individuals' online engagement regarding political-based contrasts
- Partisan: a dimension representing individuals' opinions about the usefulness of vaccines. This dimension was selected as an indicator of what is known in Social Psychology as "Partisan Bias"(*Tversky & Kahneman*, 1974; *Codagnone et al*, 2018) which leads people to align their opinion to the one of the social group in which they recognize themselves regarding contrasts between opposing views on an issue of public interest coming from different social groups.
- Political: a dimension representing individuals' support for parties-backed social battles
- Civil: a dimension representing individuals' believes on civil rights or liberties; in this particular case, on abortion.
- Party: a dimension representing individuals' preferences in terms of political party.

For each axis, two contrasting hashtags were selected based on a search for influential tweets among US' Conservatives and Liberals. (M. *Anderson et al.*, 2018; *Gunaratne et al.*, 2019). These initial classification its intended as an idealtypical representation of the two predominant views in western civilizations regarding social facts. The hashtags, for each corresponding dimension, are:

- "BlackLivesMatter" Vs "AllLivesMatter"
- "RepublicansAreTheProblem" Vs "DemocratsAreADisaster"
- "VaccinesWork" Vs "Vaxxed"
- "WeWantVotingRights" Vs "VoteThemAllOut2022"
- "ProChoice" Vs "ProLife"
- "VoteBlue" Vs "VoteRed"

For each dimension d, 10000 tweets were downloaded such that each tweet could contain either one or the other hashtag in the domain of the axis -to exclude ambiguous results- and then all the resulting tweets were grouped by user, so that information on each user could represent its stance on each issue located by the corresponding dimension. The dataframe obtained was composed of 15927 users which had used one up to six of these hashtags in their tweets. The results were coded with a multinomial classification: for each dimension, if a user had tweeted a hashtag corresponding to a progressive stance on the issue, it was assigned a value 1, if it hadn't tweeted anything on the subject it was assigned a 0 and finally if its views were aligned with that of the conservatives idealtypical paradigm, it was assigned the value -1. The underlying idea is that society is divided into groups that compose their Habit in relation with the typical social ethos of the predominant social groups in their time and space, as already mentioned in the introduction. The different dimensions analyzed also allow to identify groups on different positions of the social environment, permitting users to have mixed representation resulting from their digital practices and thus allowing to identify groups with nuanced positions in the social space.

User	Racial	Activism	Partisan	Political	Civil	Party
12301	0	0	0	1	0	0
658473	0	0	0	1	0	0
658583	0	0	0	1	0	0
698223	0	0	0	0	0	1
727483	0	0	1	0	0	0

Table 1: first five rows of the original data set.

The data matrix was then rescaled with Non-linear Principal Component Analysis (NLPCA) with optimal scaling, so that the ordinal variables could be transformed into numerical ones and used for clustering. NLPCA is a dimensionality reduction method similar to classical PCA, with the difference that it uses non-linear transformation of the original variables in the matrix decomposition, thus allowing to treat qualitative scales by converting every category to a numeric value (*Manzi*, 2021; *Meulman et al.*, 2004). Considered that the axis are constructed to be almost orthogonal, a fact reflected by the VAF of each principal component and presented in *figure 1*, all dimensions d are retained during the variable transformation.

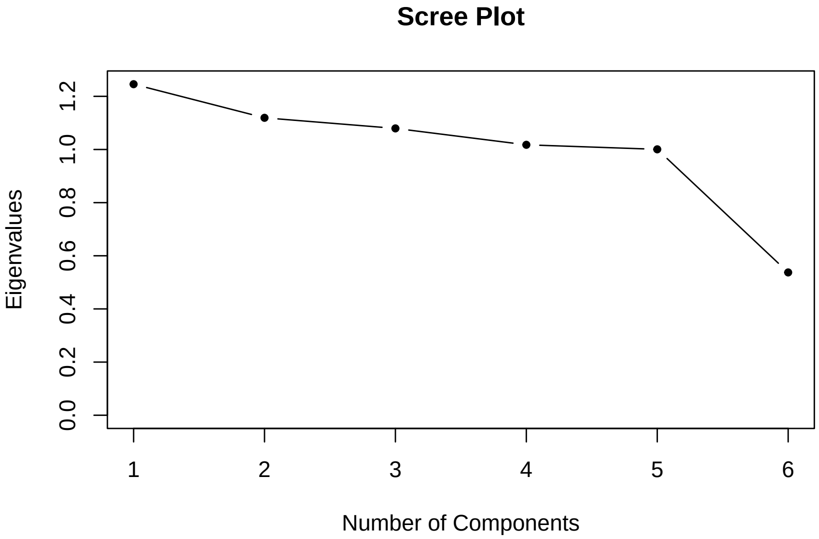

Figure 1: VAF per component

Rescaling the variables with this device require to look at the new coordinates for interpreting the resulting plot of the observations in the new d dimensional space (*Figure 2*).

Racial	Activism	Partisan	Political	Civil	Party
1.5029848	-0.2219571	-0.6196719	-1.1567127	-0.7169473	0.7778824
1.5029848	-0.2219571	-0.6196719	-1.1567127	-0.7169473	0.7778824
1.5029848	-0.2219571	-0.6196719	-1.1567127	-0.7169473	0.7778824
-1.5856812	2.9231162	-0.0641863	-1.0262169	-0.2117745	0.6226040
-0.8165267	-0.9872594	0.7518776	-0.1920446	0.4129010	0.6694000

Table 2: first five rows of the reconstructed data set.

Loadings (presented in *table 3*) are useful to map the reconstructed coordinates into the new space, giving the sense of direction; for example, a point which identify an individual of liberal beliefs about minority rights and which doesn't express an opinion on other issues will be in the top left corner of the plot, near the origin of the z axis and of the y axis, corresponding to positive values of the first three principal components.

	D1	D2	D3	D4	D5	D6
Racial	0.5024030	0.3566067	0.4872380	0.5709516	0.2056583	0.1214122
Activism	-0.0077483	0.0097387	-0.6938463	0.2307419	0.6544091	0.1921701
Partizan	-0.4823466	-0.5657494	0.4356035	-0.1114276	0.2535947	0.4251970
Political	0.6431822	-0.0736345	-0.2501212	-0.5039538	-0.2852129	0.4278063
Civil	0.2315501	0.0808674	0.2954596	-0.5827998	0.6377870	-0.3257618
Party	-0.4284745	0.8225145	-0.0293750	-0.2811945	-0.0233580	0.2437224

Table 3: loadings

The initial plots of the reconstructed observations suggest the presence of at least three macro groups.

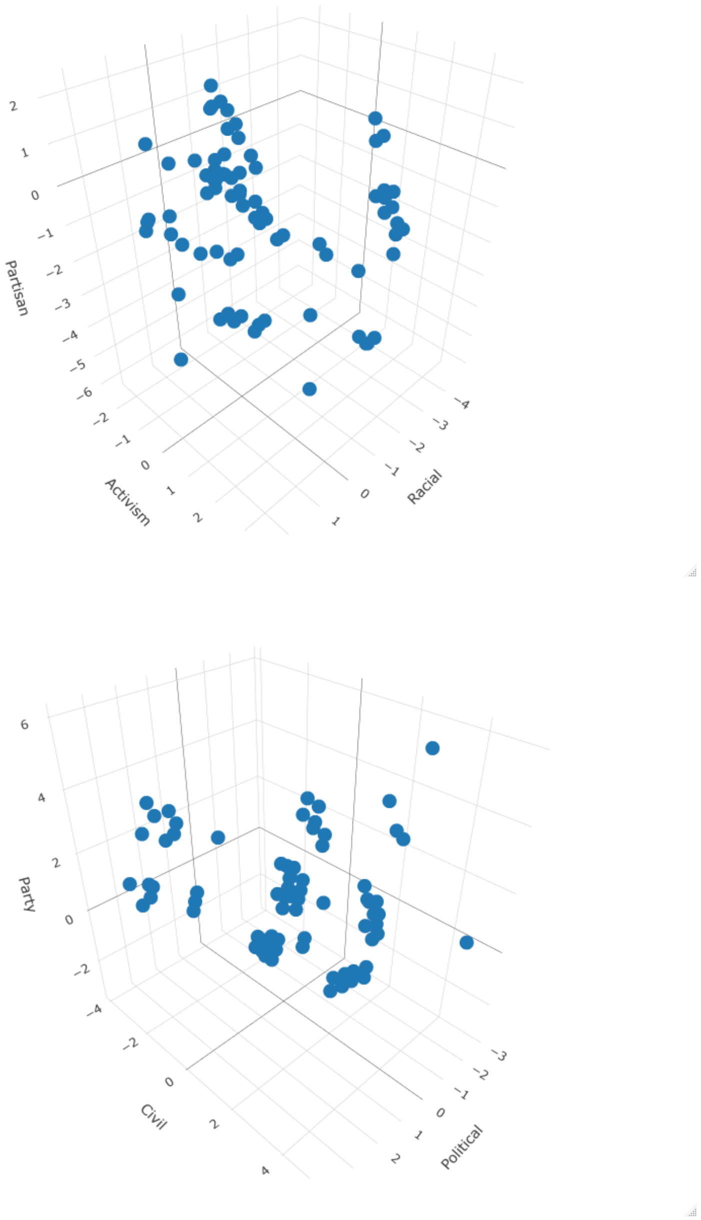

Figure 2: first first (top) and last (bottom) three dimensions of the reconstructed data set in the new space.

Gaussian mixture model and classical model-based clustering

In model-based clustering, the data $y = (y_1, ..., y_n)$ are assumed to be generated by a mixture model with density

$$f(y) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k f_k(y_i|\theta_k) \tag{1}$$

where $f_k(y_i|\theta_k)$ is a probability distribution with parameters θ_k, and τ_k is the probability of belonging to the k-th component[1]. In the multivariate Gaussian mixture model, f_k are taken to be multivariate normal distributions, parameterized by their means μ_k and covariances Σ_k:

$$f_k(y_i|\theta_k) = \phi(y_i|\mu_k, \Sigma_k) \equiv |2\pi\Sigma_k|^{-\frac{1}{2}} exp\{\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\} \tag{2}$$

where $\theta_k = (\mu_k, \Sigma_k)$.

Model-based clustering differs from other clustering methods as it doesn't involve computing the distances between observations to build the final clusters: instead, data points are assigned to groups based on the conditional probability of being generated by that component given the corresponding value of μ_k and Σ_k. The parameters of the model for each component are usually estimated through the Expectation-Maximization Algorithm (EM), based on MLE solution, to find the estimates $\widehat{\theta_k}$ (*Dempster et al.*,1977). This is a general approach to maximum likelihood for problems in which the data can be viewed as consisting of n multivariate observations (y_i, z_i), in which y_i is observed and z_i is unobserved. If the (y_i, z_i) are independent and identically distributed (iid) according to a probability distribution f with parameters θ, then the *complete-data likelihood* is

$$\mathcal{L}_{\mathcal{C}}(y, z|\theta) = \prod_{i=1}^{n} f(y_i, z_i|\theta) \tag{3}$$

where $y = (y_1, ..., y_n)$ and $z = (z_1, ..., z_n)$. The *observed data likelihood*, $\mathcal{L}_{\mathcal{O}}(y|\theta)$, can be obtained by integrating the unobserved data z out of the complete-data likelihood:

$$\begin{aligned}\mathcal{L}_{\mathcal{O}}(y, z|\theta) &= \int \mathcal{L}_{\mathcal{C}}(y, z|\theta)dz \\ &= \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \phi_k(y_i|\mu_k, \Sigma_k)\end{aligned} \tag{4}$$

The observed data likelihood in the Gaussian mixture model context is equal to (1), where $f_k(y_i|\theta_k) = \phi_k(y_i|\mu_k, \Sigma_k)$ and is generally called *Mixture likelihood.* Then the vector $z = (z_1, ..., z_n)$, where $z_i \in \{1, ..., G\}$, represents the cluster membership for observation i. Consequently the log of the complete-data likelihood (3) can be written as:

$$\ell_C(\theta_k, \tau_k, z_{i,k}|y) = \sum_{i=1}^{n} \sum_{k=1}^{G} z_{i,k} \log[\tau_k f_k(y_i|\theta_k)] \tag{5}$$

as we can assume, given that the membership of a unit i to a group k can be expressed as:

[1] *Note*: from now on, "Component" must be understood as referencing the distributions that compose the mixture, and not as a dimension of the reconstructed space obtained with NLPCA

$$z_{i,k} = \begin{cases} 1 & \text{if } y_i \text{ belongs to } k \\ 0 & \text{otherwise} \end{cases}$$

that the density of an observation y_i can be given by

$$\prod_{k=1}^{G} f_k(y_i|\theta_k)^{z_{i,k}}$$

We also assume that the Z_i are independent and identically distributed, each according to a multinomial distribution of one draw from G categories with probabilities $\tau_1, ..., \tau_G$. It is possible to use Bayes theorem to estimate the conditional probabilities that $Z_i = k|y_i$:

$$P(Z_i = k|Y_i) = \frac{P(Y_i|Z_i = k)P(Z_i = k)}{\sum_{j=1}^{G} P(Y_i|Z_i = j)P(Z_i = j)} \tag{6}$$

Note that, because $f(Z_i)$ is multinomial with values $k \in \{1, ..., G\}$ and probabilities $P(Z_i = k) = \tau_k$ for $k = 1, ..., G$, in the setting of a multivariate Gaussian mixture model we can write (6) as:

$$P(Z_i = k|Y_i) = \frac{\tau_k N(y|\mu_k, \Sigma_k)}{\sum_{j=1}^{G} \tau_j N(y|\mu_j, \Sigma_j)} \tag{7}$$

The MLE solution for the maximization of the complete-data log-likelihood (5) would yield the best parameters estimate $\widehat{\theta_k}$ for the model. However this implies estimating Z_i, which is a function of θ_k, while θ_k depends on the values of Z_i. The EM algorithm solves this problem by recursively estimating both the conditional probabilities $P(Z_i = k|Y_i)$ and the parameters θ_k in two steps: the first[2] is the Expectation step (E-step) in which, given an initial set of parameters ${\theta_k}^{(0)}$, the value $\hat{z}_{i,k}$ of $z_{i,k}$ at a maximum of (5) is the estimated conditional probability that observation i belongs to group k:

$$\hat{z}_{i,k}^{(s)} = \frac{\hat{\tau}_k^{(s-1)} f_k(y_i|\hat{\theta}_k^{(s-1)})}{\sum_{j=1}^{G} \hat{\tau}_j^{(s-1)} f_j(y_i|\hat{\theta}_j^{(s-1)})} \tag{8}$$

where subscript (s) stands for the s-th iteration of the algorithm for mixture models and subscript $(s-1)$ for the previous one. The quantity $\hat{z}_{i,k}^{(s)} = E[z_{i,k}|y_i, \theta_1, ..., \theta_G]$ for the model (1) is the conditional expectation of $z_{i,k}$ given the parameter values at the $(s-1)$th iteration and the observed data y. The second is the Maximization step (M-step), that involves maximizing (5) in terms of τ_k and θ_k with $z_{i,k}$ fixed at the values computed in the E-step, namely $\hat{z}_{i,k}$. For multivariate normal mixtures, the E-step is given by (8) with f_k replaced by ϕ_k as defined in (2), regardless of the parameterization. For the M-step, estimates of the means and probabilities have simple closed-form expressions involving the data and $\hat{z}_{i,k}$ from the E-step, namely:

$$\hat{\tau}_k^{(s)} = \frac{\hat{n}_k^{(s-1)}}{n} \; ; \; \hat{\mu}_k^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{i,k}^{(s-i)} y_i}{\hat{n}_k^{(s-1)}} \; ; \; \hat{n}_k^{(s-1)} = \sum_{i=1}^{n} \hat{z}_{i,k}^{(s-i)}.$$

Computation of the covariance estimate $\hat{\Sigma}_k^{(s)}$ depends on its parametrization. For instance, for the model with unconstrained ellipsoidal covariance, it is:

[2] *Note*: in reality the algorithm starts from the M-step to achieve computational efficiency. This means that at iteration (s) the value for $\hat{z}_{i,k}$ in the M step is actually $\hat{z}_{i,k}^{(s-1)}$ and the values of the parameters in the E-step are ${\theta_k}^{(s)}$. For this reason, the classification vector $z = (z_1, ..., z_n)$ is first computed at $(s = 0)$ based on the values of $\tau_k^{(0)}$ and is in fact required to initialize the algorithm.

$$\hat{\Sigma}_k^{(s)} = \frac{\sum_{i=1}^{n} \hat{z}_{i,k}^{(s-i)} (y_i - \hat{\mu}_k^{(s)})(y_i - \hat{\mu}_k^{(s)})^T}{\hat{n}_k^{s-1}} \tag{9}$$

At the start of each iteration (s), the log-likelihood in (4) is evaluated by replacing μ_k, Σ_k and τ_k with $\hat{\mu}_k^{(s-1)}$, $\hat{\Sigma}_k^{(s-1)}$ and $\hat{\tau}_k^{(s-1)}$. At the end of the iteration (s), the values for $\hat{\theta}_k$ in (4) are evaluated again replacing the estimates at $(s-1)$ with those of the current iteration. The algorithm stops if $\ell^{(s)}(\theta|Y_1, ..., Y_n) - \ell^{(s-1)}(\theta|Y_1, ..., Y_n) < \epsilon$, or if the maximum number of iterations is achieved. A typical value for the threshold is $\epsilon = 10^{-5}$.

The EM algorithm is widely used in model based clustering with good results, but can fail to converge, instead diverging to a point of infinite likelihood. This is because, as $\mu_k \to y_i$ and $|\Sigma_k| \to 0$ for any observation i and mixture component k, i.e. as the component mean approaches the observation and the component covariance becomes singular, then the likelihood for that observation becomes infinite and hence so does the whole mixture likelihood (4). Thus in general there is at least one path in parameter space per observation along which the likelihood tends to infinity (*Titterington et al.*, 1985). In practice, this behavior is due to singularity in the covariance estimate, and doesn't cause as much problem in terms of estimation as of regarding model selection. To overcome this problem, Fraley and Raftery propose a solution based on a Bayesian regularization of the objective function (4) for the EM and replacing its MLE estimate with a Maximum a Posteriori (MAP) estimate obtained through such regularization (*Fraley and Raftery*, 2005). The analysis carried out in this paper applies their method for multivariate normal mixture models, described in the next section.

Bayesian regularization for multivariate Gaussian mixture models

The procedure proposed by Fraley and Raftery involves placing a prior distribution on the parameters that eliminates failure due to singularity, while having little effect on stable results obtainable without a prior. The Bayesian predictive density for the data is assumed to be of the form

$$\mathcal{L}(Y|\tau_k, \mu_k, \Sigma_k)\mathcal{P}(\tau_k, \mu_k, \Sigma_k|\xi) \tag{10}$$

where $\mathcal{L}$ is the mixture likelihood as in (4):

$$\mathcal{L}(Y|\tau_k, \mu_k, \Sigma_k) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k \phi(y_i|\mu_k, \Sigma_k) = \prod_{i=1}^{n} \sum_{k=1}^{G} \tau_k |2\pi\Sigma_k|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(y_i - \mu_k)^T \Sigma_k^{-1}(y_i - \mu_k)\} \tag{11}$$

and $\mathcal{P}$ is a prior distribution on the parameters τ_k, μ_k and Σ_k, which includes other parameters denoted by ξ. The objective is to find the MAP estimate for the mixture parameters. Regarding the choice of priors for θ_k , it is assumed that:

- the mixture probabilities τ_k are uniformly distributed on the G-simplex.
- each vector mean μ_k is normally distributed (conditional on the covariance matrix)

$$\mathcal{P}(\mu_k|\Sigma_k) \sim \mathcal{N}(\mu_p, \Sigma_p/\kappa_p) \propto |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{\kappa_p}{2} trace[(\mu_k - \mu_p)^T \Sigma^{-1}(\mu_k - \mu_p)]\} \tag{12}$$

- the prior distribution for each covariance matrix Σ_k is an Inverse-Wishart

$$\mathcal{P}(\Sigma_k) \sim inverseWishart(\nu_p, \Lambda_p) \propto |\Sigma_k|^{-\frac{\nu_p+d+1}{2}} \exp\{-\frac{1}{2} trace[\Sigma_k^{-1}\Lambda_p)\} \tag{13}$$

where d is the number of dimensions and the subscript p indicates a prior hyperparameter. These are the *mean*, *shrinkage* and *degrees of freedom*, respectively μ_p , κ_p and ν_p while the hyperparameter Λ_p is the *scale* matrix of the inverse-Wishart prior. The joint prior is a normal-inverse-Wishart

$$\begin{aligned}
\mathcal{P}(\theta|\xi) &\sim Normal - inverseWishart(\mu_p, \kappa_p, \nu_p, \Lambda_p) \\
&\propto |\Sigma|^{-\frac{\nu_p+d+2}{2}} \exp\{-\frac{1}{2} trace(\Lambda_p^{-1}\Sigma^{-1})\} \exp\{-\frac{\kappa_p}{2}(\mu - \mu_p)^T \Sigma^{-1}(\mu - \mu_p)\} \\
&= |\Sigma|^{-\frac{\nu_p+d+2}{2}} \exp\{-\frac{1}{2} trace(\Lambda_p^{-1}\Sigma^{-1})\} \exp\{-\frac{\kappa_p}{2} trace[\Sigma^{-1}(\mu - \mu_p)(\mu - \mu_p)^T]\}
\end{aligned} \tag{14}$$

as the independent prior over the mixture proportions is constant and therefore τ disappears in the approximation. This is a conjugate prior for a multivariate normal distribution, because the posterior can be also expressed as a product between a normal distribution and an inverse-Wishart.

Model characterization for multivariate mixtures and posterior parameters

The choice of the hyperparameters and of the characteristics of the covariance matrix is guided by the objective of placing as little prior information on the data as possible, as well as not imposing any constraint on the Σ_k s. To do so, it was chosen an ellipsoidal model for the covariance matrices, i.e., the contour of the component densities are assumed to be ellipsoidal, as is the case for multivariate normal densities, while their volumes, shapes and orientations were allowed to vary across all components. The hyperparameters ξ are assumed to be equal across all components; following Fraley and Raftery's (2005) empirical results for the characterization of non-informative priors for the model, their values are:

- μ_p : the mean vector of the data
- κ_p : .01
- ν_p : $d + 2 = 8$
- Λ_p : $\frac{var(data)}{G^{2/d}}$

With this parameterization, (14) becomes

$$\begin{aligned}
\mathcal{P}(\theta|\xi) &\sim Normal - inverseWishart(\mu_p, \kappa_p, \nu_p, \Lambda_p) \\
&\propto \prod_{k=1}^{G} |\Sigma_k|^{-\frac{\nu_p+d+2}{2}} \exp\{-\frac{1}{2} trace(\Lambda_p^{-1}\Sigma_k^{-1})\} \exp\{-\frac{\kappa_p}{2}(\mu_k - \mu_p)^T \Sigma_k^{-1}(\mu_k - \mu_p)\} \\
&= \prod_{k=1}^{G} |\Sigma_k|^{-\frac{\nu_p+d+2}{2}} \exp\{-\frac{1}{2} trace(\Lambda_p^{-1}\Sigma_k^{-1})\} \exp\{-\frac{\kappa_p}{2} trace[\Sigma_k^{-1}(\mu_k - \mu_p)(\mu_k - \mu_p)^T]\}
\end{aligned} \tag{15}$$

Then, the posterior estimators for the mean and variance that maximize the expected complete-data log-likelihood (5) in the M-step of the EM algorithm become:

$$\hat{\mu}_k = \frac{n_k \bar{y}_k + \kappa_p \mu_p}{n_k + \kappa_p}$$

$$\hat{\Sigma}_k = \frac{\Lambda_p + \frac{\kappa_p n_k}{(n_k+\kappa_p)}(\bar{y}_k - \mu_p)(\bar{y}_k - \mu_p)^T + W_k}{\nu_p + n_k + d + 2} \tag{16}$$

where $z_{i,k}$ is the conditional probability that observation i belongs to the k-th component, $n_k \equiv \sum_{i=1}^n z_{i,k}$, $\bar{y}_k \equiv \sum_{i=1}^n \frac{z_{i,k} y_i}{n_k}$ and $W_k \equiv \sum_{i=1}^n z_{i,k}(y_i - \bar{y}_k)(y_i - \bar{y}_k)^T$. See the appendix of (*Fraley and Raftery*, 2005) for derivations.

Model evaluation and results

Given the setting discussed above, 50 possible models corresponding to $G = 1, ..., 50$ were evaluated based on their Bayesian Information Criterion (BIC) given by

$$BIC_{\mathcal{M}} = 2loglik_{\mathcal{M}}(y, \theta^*) - df_{\mathcal{M}} log(n) \tag{17}$$

where $loglik_{\mathcal{M}}(y, \theta^*)$ is the log-likelihood evaluated at the MAP for the model $\mathcal{M}$ and the data, n is the number of observations in the data and $df_{\mathcal{M}}$ is the degrees of freedom for the model $\mathcal{M}$, corresponding to $df_{\mathcal{M}} = G_{\mathcal{M}}(\frac{(d \times d - d)}{2} + 2d + 1)$.

In a normal model based setting, any component with fewer than d points will tend to have a singular covariance matrix, and hence produce an infinite likelihood, even if there is a true cluster with fewer than d points. Thus the singularities might lead to incorrect model specification, as the algorithm doesn't consider these solutions. The Bayesian regularization discussed above resolves this problem by allowing the likelihood to increase smoothly rather than jumping to infinity because, when a proper prior is defined, there are generally no paths along the parameter space in which the posterior density tends to infinity.

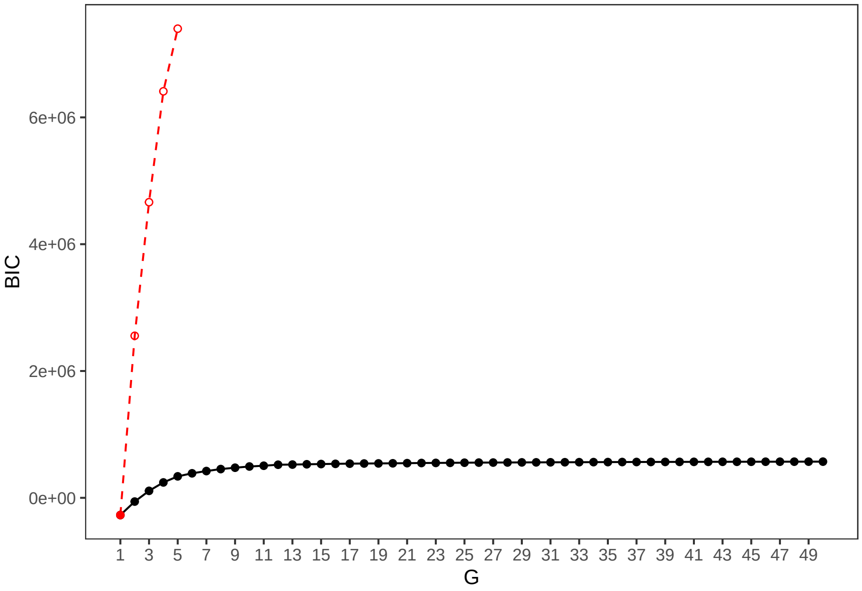

Figure 3: comparison of the two clustering methods, with and without Bayesian regularization.

Figure 3 illustrates this behaviour[3] for the algorithm comparing BIC from a model with no regularization versus the one discussed above. The classical model-based clustering, represented by the full red point, selects only the model with $G = 1$; the model with relaxed assumptions allowing spurious solutions, represented by the dashed red line, quickly jumps to infinity as $G \to 6$ while the regularized Bayesian model steadily increases for the first 9 components and then, after it reaches $G = 12$ corresponding to $BIC = 518380.23$, remains largely unchanged with small improvements. Following these results it was decided to choose $G = 9$, as the log-likelihood doesn't improve much after that value for G and the model was kept parsimonious.

The results for the chosen model are showed in *figure 4*.

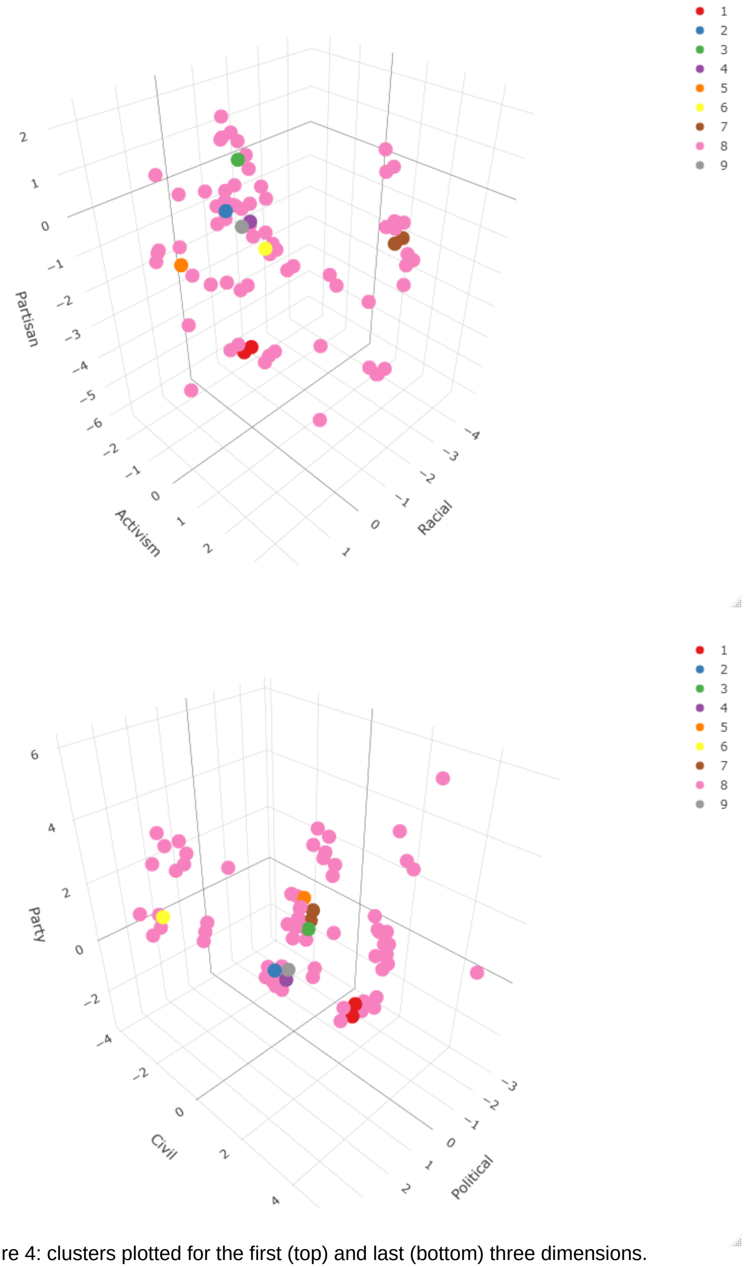

Figure 4: clusters plotted for the first (top) and last (bottom) three dimensions.

From these plots it is evident that most of the variability in the data is due to the large variance of the

[3] *Note:* the algorithm constraints were relaxed to show spurious solutions that instead would be automatically cut off to expose this difference in the two methods of estimation.

8th group, while all other components have much more concentrated densities. Note that there are many repetitions in the data, so despite what the visual inspection may suggest, group 8 is not the largest, as confirmed by the evaluated τ_k :

tau.1	tau.2	tau.3	tau.4	tau.5	tau.6	tau.7	tau.8	tau.9
0.1536385	0.2468764	0.1256985	0.0924845	0.2033026	0.027626	0.0384881	0.062912	0.0489734

Table 4: mixture proportions.

in fact, the observations per group are :

Group.1	Group.2	Group.3	Group.4	Group.5	Group.6	Group.7	Group.8	Group.9
2447	3932	2002	1473	3238	440	613	1002	780

Table 5: observations per group.

Estimates θ^*_k were output after 50 iterations of the EM algorithm, although after the first ten iterations values of the parameters did not change for much of the components, and after twenty reached convergence for all of them. This can be inspected watching the traces of the parameters, for example the mean vector μ_1, at the first and last five iterations of the algorithm.

Racial	Activism	Partisan	Political	Civil	Party
-0.00811	0.03328	-0.01999	-0.00283	0.01133	-0.02817
-0.11063	0.08211	-0.14884	-0.12473	0.08572	-0.17046
-0.11047	0.44638	0.05592	0.08025	0.07657	-0.24970
-0.18586	0.34838	-0.17266	-0.00385	-0.11590	-0.76129
-0.17480	0.24044	-0.22037	-0.05940	-0.20701	-0.90249
...	...	...	...	...	...
-0.19603	0.21568	-0.15264	-0.09824	-0.31646	-1.00708
-0.19603	0.21568	-0.15264	-0.09824	-0.31646	-1.00708
-0.19603	0.21568	-0.15264	-0.09824	-0.31646	-1.00708
-0.19603	0.21568	-0.15264	-0.09824	-0.31646	-1.00708
-0.19603	0.21568	-0.15264	-0.09824	-0.31646	-1.00708

Table 6: first and last five iterations of the EM algorithm for the mean of the first component.

Or by plotting μ's convergence with respect to one of the dimensions, say *"Activism"*:

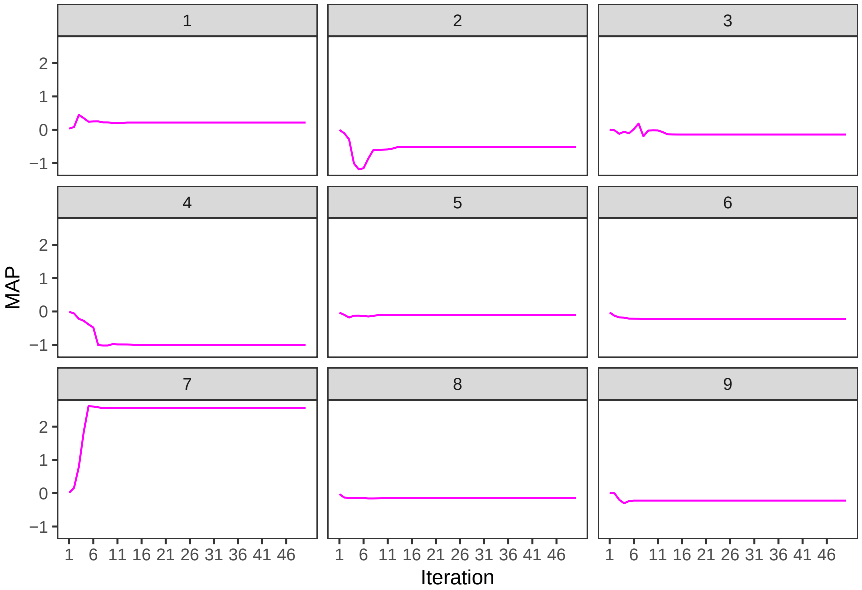

Figure 5: traces of the mean of every component for the second dimension (Activism).

This diagnostics can be repeated for all estimates, verifying the convergence and the characteristics in terms of volume, shape and orientation also for each Σ_k.

As for the interpretation of the results, it's useful to start by looking at the components' mean vectors

	Group.1	Group.2	Group.3	Group.4	Group.5	Group.6	Group.7	Group.8	Group.9
Racial	1.50298	-0.81652	-0.26444	-0.31791	0.85489	0.02629	-0.98924	-1.60207	-0.04813
Activism	-0.22196	-0.98726	-0.02189	-0.00912	0.57259	-0.11026	-0.35546	2.92494	-0.04638
Partisan	-0.61967	0.75188	-0.21548	-0.37976	0.72867	-3.84225	-1.38802	-0.06043	-0.26494
Political	-1.15671	-0.19204	0.18112	-0.81981	1.01665	1.32067	2.54456	-1.01379	0.01739
Civil	-0.71694	0.41290	-0.15664	0.18637	0.23790	3.51169	-2.87590	-0.20657	-0.22521
Party	0.77788	0.66940	-1.35856	-0.98663	-0.28140	0.68001	0.81782	0.60277	-1.09708

Table 7: groups' mean vectors.

then, by comparing the coordinates of the data points in the new space created by the recomposition of the original categorical data set with their original encoding, it's possible to deduce the characteristic of the group examined. For example, the first three observations in *table 2*, corresponding to individuals which have expressed a support for liberal political battles but didn't engage in other practices will be found in group 5. Another way is to look at the classification values $z^*_{i,k}$ output by the algorithm. Following this procedure, data show that:

- The first component represents individual which have showed a liberal alignment regarding political based contrasts, but took no further action on other issues. This group can be interpreted as holding an anti-conservative ethos, but not necessarily pro leftist, as some of them holds typical conservative positions regarding party-backed social issues (the idea that all politicians are corrupt or inefficient)
- The second component contain individuals that have expressed liberal sentiments regarding issues of racial equality. Is the second most numerous community and it's very cohesive.
- Group 3 represents people who advocate for the use of vaccines. It's the most numerous, and their member are likely of centrist or democratic political extraction.
- The fourth group is instead composed of individuals that like the idea of getting rid of the state ingerence, a retoric typically sustained by conservative parties. A few of them also support the republican-backed hashtag campaign "DemocratsAreADisaster", supporting the idea that this group identifies a populist or right-wing ethos.
- Component 5 has an opposite interpretation with respect to the previous one: this group identifies liberal democrats, people that showed support for democratic-backed social battles for voting rights, although didn't express other forms of liberal practices.
- The sixth identified community, representing true conservatives, is much smaller than the others: its members hold conservatives views regarding civil rights and are likely driven by a traditionalist ethos.
- The seventh group is composed of individuals that express their preference for the democratic party through their digital practices. However, many of them also show discontent toward politicians and question the effectiveness of vaccines; this leads to think that this group identifies the democratic party's popular base.
- As showed by the clustering, component 8 is the most heterogeneous. It includes many subgroups,as well as a couple major ones, namely people that holds a progressive stance regarding civil rights and moderate conservatives. As for the subgroups, it's possible to identify left-wing supporters, engaged democrats, right wing extremists and fundamentalist conservatives. This community is most probably a container for the minoritarian components of the social space.
- Finally group 9, opposed to the third, contain individuals that are skeptical of the vaccines and identifies non-political conservatives.

A final consideration can be made about the conditional probabilities $Z^*_{i,k}$: if the clusters are well specified, the conditional probabilities can be interpreted as indicators of the relative importance of a digital practice in determining the behaviour and ethnomethods of the individual, as they can measure in terms of probability the connection between practices and habit.

Conclusions

The method implemented constitutes a clustering procedure that can be applied to ordinal data, after NLPCA. It is robust to singularities in the covariance matrices of the components, thanks to the Bayesian regularization: this guarantees that components formed by identical observations -quite common in a context of ordinal data- that will have an estimate of their mean equal to the observations' value, won't be overlooked by the clustering algorithm thus making possible to specify the correct model. The results then allow to make valid inference about the groups' characteristics in the social space. Model based clustering is in general particularly suited for analysis of these kind, in which deterministic clustering methods would have no meaning. In fact classical hierarchical clustering works with distances between observations, a measure which doesn't convey any information when working with observations encoded by the researcher. Instead, by looking at the distributions of the data, it is assured that clusters are built based on unobserved latent variables, identified, in this particular case, with the groups' habits. Absolute distances between observations still don't hold any interpretation, however the variability inside the clusters can be interpreted as a measure of the communities' cohesion in terms of practices. The prior specification, based on the data, also allow to have a fast convergence of the EM algorithm even when it is non-informative like in this case.

Further comments

This paper provides a solid methodology to make inference about social communities based on individuals' indicators of socio-political stance. However, as for the data used in this example, it is not free of inaccuracies. In particular, inferential claims on the data are not very interesting, since the communities identified in almost all cases are composed based on the distribution of just one of the dimensions examined, leaving community 8 as the receptacle of the very information it was intended to find, but without possibly discerning any other detail because of the non-informative distances used for the model. This general problem is caused by two main issues: the sparsity of the original matrix and the lack of a measure of meaningful relative distance between observations. Solutions to both of these difficulties can be addressed in multiple ways: Data collection can be done assuring that most of the observations contain values for at least three of the original dimensions; other scales and type of indicators can be considered so that a dense matrix can be provided to the initial decomposition, as well as providing redundant information through other, non orthogonal dimensions that can be later reduced through NLPCA . However, this could still lead to relative distances that are not meaningful for inference; this may be solved by a transformation carried out through a Recurrent Neural Network (RNN) rather than by NLPCA: the resulting vectors are meaningful in terms of distances from one another (Zhu, 2021). It's also possible to address this problem at modelling level: by assuming a multinomial mixture model and choosing a suitable prior on the model parameters, for example a Dirichlet distribution, it would be possible to maintain the original encoding without the need for any transformation of the data, provided that it can convey a relative measure of distance.

References

- Anderson, M., Toor, S., Rainie, L. & Smith, A.(2018). *An analysis of #BlackLivesMatter and other Twitter hashtags related to political or social issues.* Pew Research Center.

- Bourdieu, P. (1979). *La distinction.* Les éditions de minuit. Paris.

- Bouveyron, C., Celeux, G., Murphy, T., & Raftery, A.E. (2019). *Model-Based Clustering and Classification for Data Science: With Applications in R.* Cambridge University Press. Cambridge.

- Castells, M. (1996) *The rise of the network society.* Blackwell Publishers. Malden.

- Castells, M. (2009) *The power of identity; The Information Age: Economy, Society, and Culture Volume II.* Wiley-Blackwell. Berkeley. Berkeley.

- Chen, T., Zhang, N.L., Liu, T.K., Poon,M. & Wang, Y. (2012). *Model-based multidimensional clustering of categorical data.* Artificial Intelligence Volume 176, Issue 1, January 2012, pp.2246-2269.

- Codagnone, C., Bogliancino, F. & Veltri, G.A. (2018) *Scienza in vendita: Incertezza, interessi e valori nelle politiche pubbliche.* Egea. Milan.

- Couldry, N. (2012). *Media, Society, World. Social theory and digital media practice*, Polity Press ltd. Cambridge.

- Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). *Maximum Likelihood from Incomplete Data via the EM Algorithm*, Journal of the Royal Statistical Society. Series B (Methodological), Vol. 39, No. 1. (1977), pp. 1-38

- Fraley, C. & Raftery, A.E. (2005). *Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering.* Technical Report No. 486 Department of Statistics, University of Washington, Seattle.

- Garfinkel, H. (1967). *Studies in Ethnomethodology.* Prentice-Hall. Englewood Cliffs-New York.

- Goffman, E. (1959). *The presentation of Self in Everyday Life.* Bantam Doubleday Dell Publishing Group.

- Gunaratne, K., Coomes, E.A. & Haghbayan, H. (2019). *Temporal trends in anti-vaccine discourse on Twitter.* Vaccine, Volume 37, Issue 35, pp. 4867-4871.

- Lau, J.W. & Green, P. J. (2007). *Bayesian Model-Based Clustering Procedures.* Journal of Computational and Graphical Statistics. 16:3, pp. 526-558.

- Lyon, D. (2003). *Surveillance as Social Sorting, in Surveillance as Social Sorting: Privacy, Risk and Digital Discrimination*, D. Lyon Ed. Routledge. London.

- Manzi, G. (2021). *Lectures at University of Milan, Data Science and Economics, Advanced Multivariate Analysis.* (September-November 2021).

- Mead, G. H. (1934). *Mind, Self, and Society* (C. W. Morris, Ed.). University of Chicago Press.

- Meulman, J., Van der Kooij, A.J. & Heiser, W. (2004) *Principal Component Analysis with Nonlinear Optimal Scaling Transformations for Ordinal and Nominal Data,* Handbook of Quantitative Methods in the Social Sciences, pp.49-70, D. Kaplan.

- Schütz, A. (1944). *The Stranger: An Essay in Social Psychology.* The American Journal of Sociology, Vol. 49, No. 6 (May,1944), pp. 499-507.

- Tajfel, H. (1978). *Differentiation between Social Groups: Interindividual Behaviour and Intergroup Behaviour.* Academic Press London, New York and San Francisco.

- Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* Wiley. University of California.

- Tversky, A. & Kahneman, D. (1974). J*udgment under uncertainty: Heuristics and biases.* Science, 185(4157), pp. 1124 – 1131.

- Zhu, C. (2021). *Machine Reading Comprehension, Chapter 3: Deep learning in natural language processing.* Elsevier.