Fast Bayesian clustering and model selection for longitudinal data mixtures

Marco Corneli, Elena Erosheva

Fast Bayesian clustering and model selection for longitudinal data mixtures

M. Corneli[1] and E. Erosheva[2,3]

[1] *Université Côte d'Azur, Center of Modeling, Simulation & Interaction, France*
[2] *Department of Statistics, School of Social Work, and the Center for Statistics and the Social Sciences, University of Washington, Seattle, WA, 98195, USA*
[3] *Université Côte d'Azur, Laboratoire J. A. Dieudonné, CNRS, Nice, 06100, France*

Abstract

The clustering of longitudinal data from a Bayesian perspective is considered, with particular attention to the selection of the number of components. Instead of using asymptotic criteria (e.g. BIC), we propose to directly maximize an exact quantity based on conjugated prior distributions of the model parameters. The prior parameters are estimated by gradient descent, via automatic differentiation. Using simulated data, we demonstrate that, in terms of accuracy of the obtained clustering, our approach is comparable to two frequentist approaches commonly used in this setting, and it outperforms them in selecting the actual number of clusters.

1 Framework

We consider longitudinal data with D measurements for N individuals. Thus, a vector $y_i \in \mathbb{R}^D$ keeps track of the measurements for the i-th individual. In the following, it is assumed that measurements are taken at the same times across individuals and that the number of measurements for each individual is always equal to D. This assumptions plays a crucial role in Section 3. The j-th measurement for the i-th individual (y_{ij}) is taken at time t_j, with $j \leq D$. The following generative model is adopted

$$y_{ij} = \phi(t_{ij})^T \beta_{z_i} + \sigma \epsilon_{ij}, \quad \forall i \leq N, j \leq D, \tag{1}$$

where z_i is a *latent* random variable labeling the group (*cluster*) of the i-th individual, $\phi(\cdot) : \mathbb{R} \to \mathbb{R}^K$ is a user defined *feature map* (e.g. identity, polynomial,

etc.), $\beta_1, \dots, \beta_Q$ are parameters in $\mathbb{R}^K$ and σ is the scalar standard deviation of the noise term. The residuals ϵ_{ij} are all independent and Gaussian distributed $\mathcal{N}(0,1)$. The following more compact notation will also be employed

$$y_i = \Phi\beta_{z_i} + \sigma\epsilon_i, \tag{2}$$

where $\Phi : \mathbb{R}^D \to \mathbb{R}^{D\times K}$ is the *design matrix*, whose j-th row is $\phi(t_{ij})^T$ and $\epsilon_i \in \mathbb{R}^D$ follows a multivariate isotropic Gaussian distribution. In the following, $y_1, \dots, y_N$ are assumed independent.

Remark 1. *Notice that, as long as the number of measurements* and *the measurement times are the same for each individual,* Φ *does not depend either on* i *or* z_i*.*

Remark 2. *The generative model detailed so far can be extended straightforward by assuming that the standard deviation* σ *also depends on* z_i*. All the results reported in this paper will still be valid.*

1.1 A frequentist approach and some related drawbacks

Assuming that $z_1, \dots, z_N$ are i.i.d. random variables such that

$$\mathbb{P}(z_i = q) = \pi_q, \qquad \forall q \leq Q \tag{3}$$

where $\pi_q \in [0,1]$ and $\sum_q \pi_q = 1$ and thanks to Remark 1, it is easy to see that $y_1, \dots, y_N$ are i.i.d. following the mixture distribution

$$p(y_i|\theta, Q) = \sum_{q=1}^{Q} \pi_q g(y_i; \Phi\beta_q, \sigma^2 I_D), \tag{4}$$

where z_i was integrated out, $\theta := \{\beta_q, \pi_q, \sigma^2\}_{q\leq Q}$ denotes the set of the model parameters and $g(\cdot; \mu, \Sigma)$ denotes the pdf of a multivariate Gaussian distribution with mean μ and covariance matrix Σ. A standard approach to estimate the parameters of the model described so far (included Q) would consist into (see for instance Muthén and Shedden, 1999; Muthén and Asparouhov, 2008):

1. maximizing the log-likelihood $\sum_{i=1}^N \log p(y_i|\theta, Q)$ with respect to θ, for instance via an EM algorithm. Once obtained $\hat{\theta}^{ML}$, the posterior distribution $p(z_i|y_i, \hat{\theta}^{ML}, Q)$ could be computed for all i and used to cluster the observations. Then,

2. the number of components Q might be estimated via some model selection criterion (e.g. AIC or BIC).

There are drawbacks to using AIC/BIC for selecting the number of latent mixture components. For example, BIC is an *asymptotic* criterion needing a large number of observations to be given and some relevant assumptions to be fulfilled by the data that may not be straightforward in case of longitudinal observations. In addition, BIC was shown to possibly overestimate of the number of components in mixture models (Biernacki et al., 2000; Baudry et al., 2010).

Under a Bayesian model-based clustering framework, we propose a model selection criterion that is both more conservative than BIC and exact (i.e., not based on asymptotic theory). A natural choice is the Integrated Classification Likelihood (ICL, Biernacki et al., 2000). See Appendix A for a formal definition of ICL and the proof that ICL is a lower bound of BIC.[1]

In Section 2, we propose a Bayesian framework to perform longitudinal data clustering and model selection based on Eq. (2). Then, is Section 3 we detail an original inference procedure *not* sampling based and prove some results allowing us to dramatically speed up the estimation algorithm. Finally, in Section 4 we report simulation study results demonstrating that our method outperforms some frequentist alternative methods. A final section concludes the paper

2 A Bayesian perspective

In this section, we detail a Bayesian framework allowing us to i) cluster the observations $y_1, \ldots, y_N$ in Q groups and ii) select Q in a non asymptotic framework. The target probability distribution, that we would like to maximize with respect to the pair (Z, Q) is

$$p(Z, Q|Y) = \int p(Z, Q|Y, \theta)p(\theta)d\theta \tag{5}$$

where $Z := (z_1, \ldots, z_N)$, $Y := (y_1, \ldots, y_N)$ and the model parameters θ are seen as random variables and integrated out. Note that, from a full Bayesian perspective, the number of clusters Q is also view as a random variable in the above equation. In order to develop an estimation algorithm that is reasonably fast, we choose not to implement MCMC algorithms to simulate the above posterior distribution and present in the following an alternative strategy. Thanks to the Bayes rule it holds that

$$p(Z, Q|Y) = \frac{p(Y, Z|Q)p(Q)}{p(Y)}.$$

[1] As another alternative to BIC we also cite the slope heuristic (Birgé and Massart, 2007), a criterion whose penalty has a multiplicative factor that can be estimated from the data. However, this method requires to compute several log-likelihoods (one for each candidate Q) and will not be considered in this paper.

Since the denominator does not depend on (Z, Q), if we further assume that the prior distribution of Q is uniform ($p(Q) \propto 1$) it holds that

$$\arg\max_{(Z,Q)} p(Z, Q|Y) = \arg\max_{(Z,Q)} p(Y, Z|Q) \tag{6}$$

2.1 A closed form complete data integrated log-likelihood

The complete data integrated log-likelihood on the right hand side of Eq. (6) is

$$p(Y, Z|Q) = \int p(Y, Z|\theta, Q)p(\theta|Q)d\theta, \tag{7}$$

where we recall that $\theta := \{\beta_q, \pi_q, \sigma^2\}_q$. Note that this quantity is the one that the ICL criterion seeks to approximate (see Appendix A). To keep the notation uncluttered, we denote $\beta := \{\beta_q\}_q$ and $\pi := \{\pi_q\}_q$. Assuming that the prior distribution factorizes over the model parameters, namely

$$p(\beta, \sigma, \pi) = p(\beta, \sigma)p(\pi),$$

the integrated log-likelihood in Eq. (7) factorizes too

$$p(Y, Z|Q) = p(Y|Z, Q)p(Z|Q). \tag{8}$$

By adopting prior conjugated distributions for β, σ and π, the above likelihood can be expressed in a closed form. Thus, it is first assumed that, conditionally on σ^2, $\beta_1, \ldots, \beta_Q$ follow independent Gaussian prior distributions

$$\beta_q \sim \mathcal{N}\left(0, \sigma^2\eta_q I_K\right), \tag{9}$$

where $\eta_1, \ldots, \eta_q$ are positive parameters and I_K is the identity matrix of order K. The β_qs are further assumed to be independent from ϵ_i, for all i. Then, σ^2 is assumed to follow an Inverse Gamma prior distribution

$$\sigma^2 \sim \text{I}\Gamma(a, b), \tag{10}$$

where $a, b > 0$. Finally, π is assumed to follow a Dirichlet prior distribution

$$\pi \sim \text{Dir}(\alpha_1, \ldots, \alpha_Q) \tag{11}$$

where $\alpha_q > 0$ for all q. We will see how to maximize $p(Y, Z|Q)$ in Eq. (8) with respect to Z and Q in the next section.

First, we focus on the first term on the right hand side of the equality in Eq. (8), $p(Y|Z, Q)$.

2.1.1 Integrating with respect to β

By integrating out β_q in Eq. (2), we obtain the marginal conditional density of y_i

$$y_i|z_i, \sigma^2 \sim \mathcal{N}\left(0, \sigma^2\left(\eta_{z_i}\Phi\Phi^T + I_D\right)\right). \tag{12}$$

By the law of iterated expectations it follows that

$$\begin{aligned}\mathrm{Cov}(y_i, y_j|z_i, z_j, \sigma^2) &= \mathbb{E}\left[y_i y_j^T|z_i, z_j, \sigma^2\right] \\ &= \mathbb{E}\left[\Phi\beta_{z_i}\beta_{z_j}^T\Phi^T|z_i, z_j, \sigma^2\right] \\ &= \sigma^2\eta_{z_i}\Phi\Phi^T\mathbf{1}_{z_i=z_j},\end{aligned}$$

where $\mathbf{1}_A(\cdot)$ denotes the indicator function over a set A. The above equation has an important consequence: after β_q is integrated out, the random vectors y_i sharing the same cluster are no longer independent, as they were in the frequentist model. Moreover, let us denote by

$$y^{(q)} := \{y_i|i \leq N, z_i = q\}, \tag{13}$$

the q-th cluster, whose cardinality is denoted by C_q. Since all vectors in $y^{(q)}$ are Gaussian distributed, their joint conditional density can be specified as

$$Y_q|Z, \sigma \sim \mathcal{N}\left(0, \sigma^2 G_q\right), \tag{14}$$

where $Y_q \in \mathbb{R}^{DC_q}$ is a column vector obtained by concatenating all the observations in cluster $y^{(q)}$ and $G_q \in \mathbb{R}^{DC_q \times DC_q}$ is a block matrix. The blocks on the main diagonal are of the form $\left(\eta_q\Phi\Phi^T + I_D\right)$, whereas the blocks outside the main diagonal look like $\left(\eta_q\Phi\Phi^T\right)$.

Remark 3. *The assumption formulated in Eq.* (9) *can be rephrased be saying that all the y_i's in the same cluster $y^{(q)}$ are obtained as random perturbations around a same* signal. *This signal is a Gaussian process.*

Remark 4. *Depending on the nature of the feature map Φ, the generative model detailed so far can be expressed in the very same way in* kernel *terms. For instance, if ϕ is defined as the identity function, then $\Phi\Phi^T$ corresponds to the linear kernel, whose entry (j, l) is $\eta_q t_j t_l$.*

2.1.2 Integrating with respect to σ^2

In the previous section we saw that the marginal conditional density $p(Y|Z,\sigma^2,Q)$ is

$$\begin{aligned} p(Y|Z,\sigma^2,Q) &= \prod_{q=1}^{Q} p(Y_q|Z,\sigma^2) \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{DN}{2}} \prod_{q=1}^{Q} \sqrt{\det(G_q)}} \\ &\quad \times \exp\left(-\frac{1}{2\sigma^2}\sum_{q=1}^{Q} Y_q^T (G_q)^{-1} Y_q\right), \end{aligned}$$

where $G_q \in \mathbb{R}^{DC_q \times DC_q}$ is the block matrix introduced in Eq. (14). Since, Eq. (10) states that

$$p(\sigma^2|a,b) = \frac{b^a}{\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a-1} \exp\left(-\frac{b}{\sigma^2}\right) \mathbf{1}(\sigma^2)_{]0,\infty[},$$

when looking at the joint conditional density $p(Y,\sigma^2|Z,Q)$ as a function of σ^2 we recognize the pdf of an Inverse Gamma distribution $\text{I}\Gamma\left(a+\frac{DN}{2}, b+\frac{1}{2}\sum_{q=1}^{Q} Y_q^T G_q^{-1} Y_q\right)$. Therefore, σ^2 can be integrated out to obtain

$$\begin{aligned} p(Y|Z,Q) &= \frac{1}{(2\pi)^{\frac{ND}{2}} \prod_q \sqrt{\det(G_q)}} \frac{b^a}{\Gamma(a)} \\ &\quad \times \frac{\Gamma\left(\frac{DN}{2}+a\right)}{\left(b+\frac{1}{2}\sum_q Y_q^T (G_q^{-1}) Y_q\right)^{\frac{DN}{2}+a}}. \end{aligned} \tag{15}$$

2.1.3 Integrating with respect to π

The second integral on the right hand side of Eq. (8) can be computed in a similar fashion. Due to Eq. (11), the posterior distribution functions $p(\pi|Z,Q)$ is still a Dirichlet. It can easily be seen that

$$p(Z|Q) = \frac{\Gamma\left(\sum_{q=1}^{Q}\alpha_q\right)}{\prod_{q=1}^{Q}\Gamma(\alpha_q)} \frac{\prod_{q=1}^{Q}\Gamma(C_q+\alpha_q)}{\Gamma\left(N+\sum_{q=1}^{Q}\alpha_q\right)} \tag{16}$$

3 Inference

As shown in the previous section, the conjugate prior distributions on θ allowed us to obtain the integrated log-likelihood $\log p(Y,Z|Q)$ in a closed form, depending

on the hyper-parameters $\iota := \{\eta, a, b, \alpha\}$. If the number of groups Q is assumed to vary in a range $\{1, \ldots, Q_{\max}\}$, for a given Q, we aim at estimating

$$Z_Q^* := \arg\max_{Z|Q} \log p(Y, Z|Q).$$

Then, the Q leading to the highest value of $\log p(Y, Z_Q^*|Q)$ will be retained as the estimated number of clusters (see Section 2).

The estimation of Z_Q^* is challenging. The strategy that we propose relies on a greedy maximization of $\log p(Y, Z|Q)$. Similar approaches where used by Côme and Latouche (2015); Corneli et al. (2016), for instance, in the context of graph data clustering. The approach slightly changes depending on whether ι is considered as a fixed hyper-parameter or a parameter to optimize.

3.1 Optimizing with respect to parameters ι

In order to estimate Z_Q^*, for a given Q, we rely on the following two steps:

1. **Classification Step (CS).** Assume that ι is fixed to some value and a configuration $Z|Q$ is given. We compute changes in the integrated log-likelihood obtained by switching y_i from its current cluster to any other cluster, and retain the movement leading to the highest increase in the log-likelihood. All the vectors $y_1, \ldots, y_N$ are switched once in such a way. Notice that, since Q is given, if y_i is alone in its cluster, no movement is allowed.

2. **Maximization Step (MS).** Holding fixed the configuration $Z|Q$ obtained in the previous step, ι is optimized via automatic differentiation (stochastic gradient descent).

The two steps are repeated alternatively until no further increase of $\log p(Y, Z|Q)$ is possible. We stress that the algorithm described so far in not guaranteed to converge to a global optimum. Indeed both the CS (being a *greedy* step) and the MS could lead to a local maximum of the integrated log-likelihood. In order to reduce this risk, either the algorithm should be run several times (for each Q), with different initialization, or the algorithm might be provided with a "clever" initialization of Z, for instance obtained by k-means clustering. This latter solution is adopted in the experiments in Section 4.

Finally, we notice that the two-step algorithm described in this section is reminiscent of the Classification EM algorithm (C-EM, Celeux and Govaert, 1991). Variational extensions of the C-EM algorithm have recently been used for clustering and co-clustering purposes (see e.g. Bouveyron et al., 2016; Bergé et al., 2019). However, the algorithm described in this section exhibits a fundamental difference

with respect to C-EM algorithm(s). Indeed, while the latter optimizes the complete data log-likelihood with respect to model parameters, the former optimizes with respect to the prior distributions parameters, since the model parameters have been integrated away. Moreover, in the above cited works, the M step of the algorithm has a closed form, whereas our MS relies on stochastic gradient descent.

3.2 Fixed hyper-parameters ι

When the dataset is small or a strong prior knowledge about the modeled phenomenon is available, it might be advisable to use fixed values for the hyper-parameters ι. In this case, Z_Q^* can be estimated via the CS only, which is repeated until no further increase in the integrated log-likelihood is possible. Notice also that an intermediate approach will always be possible by only fixing some hyper-parameters in ι and optimizing with respect to the others.

3.3 Classification Step

Consider the observation $y_i \in \mathbb{R}^D$ and assume it currently belongs to the q-th cluster, namely $z_i = q$. The change in the integrated log-likelihood due to a switch of y_i from the q-th cluster to the l-th cluster can be computed as

$$\begin{aligned} \Delta_{ll}^{i:q\to l} :&= \log p(Y, Z^a|Q) - \log p(Y, Z^b|Q) \\ &= \log p(Y|Z^a, Q) - \log p(Y|Z^b, Q) \\ &+ \log p(Z^a|Q) - \log p(Z^a|Q), \end{aligned} \tag{17}$$

where Z^a and Z^b denote the configurations *after* and *before* the switch, respectively. As it can be seen by looking at Eqs. (15)-(16), the calculation of $\Delta_{ll}^{i:q\to l}$ basically boils down to compute i) the determinant of G_q and ii) the quadratic term $Y_q^T G_q^{-1} Y_q$, for all $q \leq Q$. We report in the following some results allowing us to speed up the calculation of these two terms.

First term. First, recall that

$$G_q = \left(\begin{array}{c|c|c|c} B_q & A_q & \dots & A_q \\ \hline A_q & B_q & \dots & A_q \\ \hline A_q & A_q & \ddots & A_q \\ \hline A_q & \dots & \dots & B_q \end{array}\right) \tag{18}$$

where $A_q = \eta_q \Phi\Phi^T \in \mathbb{R}^{D\times D}$ and $B_q = A_q + I_D$. Since the size of G_q changes whenever an observation is switched from one cluster to another, we need a fast way

to compute $\det(G_q)$[2]. Theorem 1 in Silvester (2000), whose precise formulation is reported in Appendix B (additional material), can help us. This theorem basically allows us to compute $\det(G_q)$ in two steps:

1. a first intermediate determinant (ID_q) is computed as if A_q, B_q in Eq. (18) where numbers
$$ID_q := \overline{\det}(G_q) \in \mathbb{R}^{D\times D}, \tag{19}$$
where the over line is used to differentiate this determinant from the real one, that we are actually trying to compute. Then,

2. since ID_q is itself a matrix, $\det(G_q)$ is computed as
$$\det(G_q) = \det(ID_q). \tag{20}$$

According to Theorem 1 in Silvester (2000), the above equality holds as long as all the possible matrix products between two blocks in G_q are commutative. This condition is fulfilled as stated in the following

Proposition 1. *The product between each pair of blocks of G_q is commutative.*

Proof. For simplicity, in this proof the subscript q is removed from A_q and B_q, since not needed. Clearly the products AA and BB are commutative. Moreover,

$$AB = A(A + I_D) = AA + A = (A + I_D)A = BA$$

and the proposition is proven. □

Now we can state the following Theorem

Theorem 1. *The determinant of G_q can be computed in $\mathcal{O}(D)$ as*

$$\det(G_q) = \prod_{j=1}^{D}(1 + C_q\lambda_j^{(q)}), \tag{21}$$

where $\lambda_1^{(q)}, \ldots, \lambda_D^{(q)}$ are the eigenvalues of A_q.

Proof. The proof of this Theorem relies on Lemma 1 in Appendix B (additional material), stating that ID_q in Eq. (19) is

$$ID_q = I_D + C_qA_q,$$

[2] For instance, the cost of computing $\det(G_q)$ via an LU decomposition is $\mathcal{O}(D^3N^3)$ and this approach is used by most linear algebra libraries.

which is a matrix in $\mathbb{R}^{D\times D}$. Now, since A_q is symmetric it admits a diagonal representation $A_q = Q_q\Lambda_q Q_q^T$, where $\Lambda_q \in \mathbb{R}^{D\times D}$ is a diagonal matrix whose non null entries are the eigenvalues of A_q and Q_q is an orthogonal matrix whose columns are the corresponding eigenvectors. Thus

$$\begin{aligned}\det(ID_q) &= \det(I_D + C_q Q_q \Lambda_q Q_q^T)\\ &= \det\left(Q_q(I_D + C_q\Lambda_q)Q_q^T\right)\\ &= \det\left(I_D + C_q\Lambda_q\right)\\ &= \prod_{j=1}^{D}(1 + C_q\lambda_j^{(q)}).\end{aligned}$$

□

Second term. The quadratic form

$$Y_q^T G_q^{-1} Y_q \tag{22}$$

is now considered. Adopting the notation in Eq. (18), Theorem 3 in Appendix B (additional material) states that the inverse matrix G_q^{-1} is also a diagonal block matrix

$$G_q^{-1} = \left(\begin{array}{c|c|c|c} V_q & W_q & \dots & W_q \\ \hline W_q & V_q & \dots & W_q \\ \hline W_q & W_q & \ddots & W_q \\ \hline W_q & \dots & \dots & V_q \end{array}\right)$$

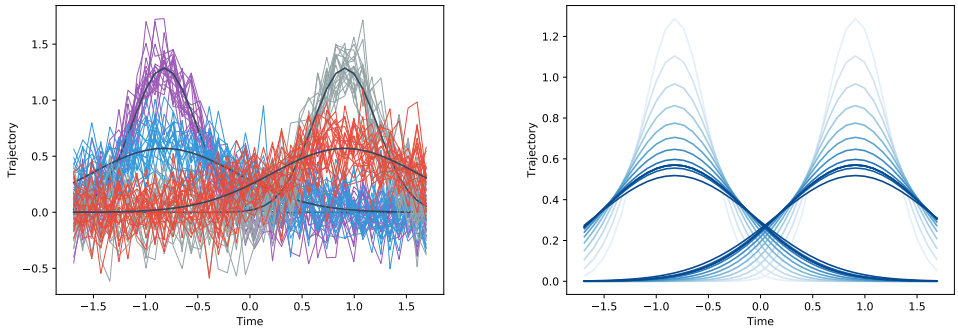
where

$$\begin{aligned} W_q &:= -(I_D + C_q A_q)^{-1} A_q,\\ V_q &:= W_q + I_D. \end{aligned} \tag{23}$$

Thus, to compute the quadratic form in Eq. (22) it is not required to invert the whole G_q but only the matrix $(I_D + C_q A_q)$. Notice that the computational cost of this operation is independent of the number of observations N[3]. Moreover

$$\begin{aligned} Y_q^T G_q^{-1} Y_q &= \sum_{i=1}^{C_q} {y_i^{(q)}}^T V_q y_i^{(q)} + \sum_{i=1}^{C_q}\sum_{\substack{j=1\\ j\neq i}}^{C_q} {y_i^{(q)}}^T W_q y_j^{(q)}\\ &= \sum_{i=1}^{C_q} {y_i^{(q)}}^T I_d y_j^{(q)} + \sum_{i=1}^{C_q}\sum_{j=1}^{C_q} {y_i^{(q)}}^T W_q y_j^{(q)}\\ &= \|Y_q\|_2^2 + \left(\sum_{i=1}^{C_q} {y_i^{(q)}}^T\right) W_q \left(\sum_{j=1}^{C_q} y_j^{(q)}\right), \end{aligned}$$

[3]For instance, relying on the Gauss-Jordan elimination, the computational cost of the inversion would be $\mathcal{O}(D^3)$, which is smaller than $\mathcal{O}(ND^3)$ required to invert G_q.



(a) (b)

Figure 1: Figure 1a shows random trajectories around four signals (case $\gamma = 0.3$). Figure 1b shows a range of curves for the four signals as γ increases up to 0.7, with darker blue lines corresonding to higher values of γ.

where $\|\cdot\|_2$ denotes the Euclidean norm and we denoted by $y_i^{(q)} \in \mathbb{R}^D$ the i-th column vector in cluster $y^{(q)}$. Since the sum of all observations in cluster $y^{(q)}$ (namely $\sum_{j=1}^{C_q} y_j^{(q)}$) can be pre-computed before and after each switch and W_q does not depend on i, the last term on the right hand side of the above equation can be computed in $\mathcal{O}(D^2)$ that can be sensibly smaller than $\mathcal{O}(N^2D^2)$ needed for a direct calculation of $Y_q^T G_q^{-1} Y_q$.

4 Experiments on simulated data

Based on a simulation study, this section compares the Bayesian approach described with two frequentist alternatives. The comparison focuses both on the accuracy of clustering for given Q and on model selection when Q is unknown. In our Bayesian framework, model selection is performed via the procedure described in Section 3. For the other two methods both BIC and ICL are computed.

The function flexmix in the eponymous R package (https://cran.r-project.org/web/packages/flexmix/index.html) fits mixtures of generalized linear models, thus it can be used to fit the generative model in Eq. (4), but with σ also depending on the cluster of the observation. That model is very similar to the frequentist version of the generative model introduced in Section 2. Since longitudinal data can be seen as high dimensional data, clustering and model selection can also be performed via the function hddc in the R package HDclassif (https://cran.r-project.org/web/packages/HDclassif/index.html) based on a latent mixture model which fits high dimensional data in group-specific subspaces (Bouveyron et al., 2007; Bergé et al., 2012).

The simulation study is based on $Q = 4$ signals being the probability density functions of the four Gaussian distributions: $\mathcal{N}(-0.83, \gamma)$, $\mathcal{N}(-0.83, 0.7)$, $\mathcal{N}(0.91, \gamma)$ and $\mathcal{N}(0.91, 0.7)$. Here, the standard deviation γ is a free parameter varying in the range $[0.3, 0.7]$. The four smooth dark curves in Figure 1a show the four signals with $\gamma = 0.3$. We use a regular grid of 40 time points between -1.5 and 1.5 and, for each time point, we sample 50 Gaussian perturbations around each signal in that point, with a standard deviation $\sigma = 0.2$. In total, $N = 200$ trajectories are sampled (colored curves in Figure 1a). Note that model selection becomes harder as γ increases up to 0.7. Eventually, as Figure 1b shows, only two distinct signals can be detected.

Clustering. First we focus on the accuracy of clustering, without considering model selection. We sampled 10 datasets of 200 trajectories, one dataset for each value of $\gamma \in [0.3, 0.7]$. Each clustering algorithm was run on each dataset provided with the actual value of $Q = 4$ and an initial value of Z obtained via k-means clustering[4]. However, note that they do not share the very same initial value of Z, since hddc (for instance) run a k-means internally. Adjusted Rand indexes (ARIs, Rand, 1971) are used to asses the obtained cluster assignments. We recall that, in clustering analysis, when comparing two label vectors Z_1 and Z_2, the ARI takes values in $[0, 1]$, where 1 means $Z_1 = Z_2$ (up to label switching) and 0 means that Z_1 and Z_2 are as far as two independent, purely random cluster assignments. Results can be seen in Figure 2. The orange curve refers to the k-means used to initialize our clustering algorithm (blue curve, henceforth called "Bayes"). As it can be seen, Bayes can always slightly improve the ARI obtained by k-means and it works particularly well for values of γ in $[0.5, 0.6]$. On the contrary, hddc and flexmix are slightly stronger for higher values of $\gamma > 0.6$. However the differences in the ARIs of the three algorithms on the whole range $[0.3, 0.7]$ do not seem significant and they might be attributed to the initialization. We feel that the three algorithms have very similar performances in terms of clustering. Some technical details about how Bayes was set. A polynomial kernel of order 6 was employed (same settings for flexmix). See in Figure 3 the MAP estimates of the means of the predictive distributions on time (case $\gamma = 0.41$). The initial values of the parameters ι were set as $\eta_q = 1$, for all q and $a = b = 1$. In our simulated experiments, we found that different initial values for these parameters did not impact on the final result. The Dirichlet parameter α is less trivial to set. This point will be discussed in next section. Regarding the Maximization Step (Section 3.1), maybe due to a lack of regularity of the integrated log-likelihood, we needed to fix the learning rate of the

[4]Equivalent results to those reported in this section were obtained when providing each algorithm with ten random initializations and retaining the estimates corresponding to the highest log-likelihood (not reported).

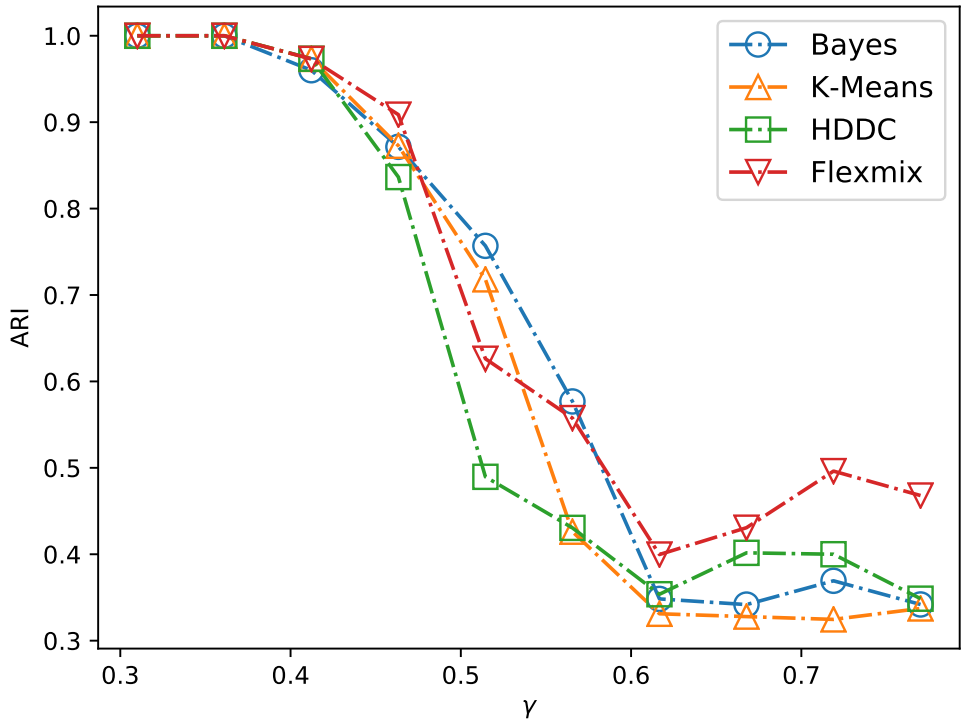


Figure 2: Adjusted Rand indexes obtained by the clustering algorithms for different values of γ.

SGD algorithm to the small value of 2e-04 and we stopped the last MS after 100 epochs. An example of the prior parameters estimates after optimization can be seen in Table 1.

Table 1: Final values of the prior parameters after optimization ($\gamma = 0.3$).

Parameter	Initial value	Final value
η	$(1.0, 1.0, 1.0, 1.0)$	$(0.99, 0.95, 0.96, 0.97)$
a	1.0	1.14
b	1.0	0.74
α	100	92.22

Model Selection. Concerning the choice of Q, the three clustering algorithms were run again on each simulated dataset, testing Q in $\{1, 2, 3, 4, 5, 6\}$. Being other settings as before, the value of Q leading to the highest value of the model

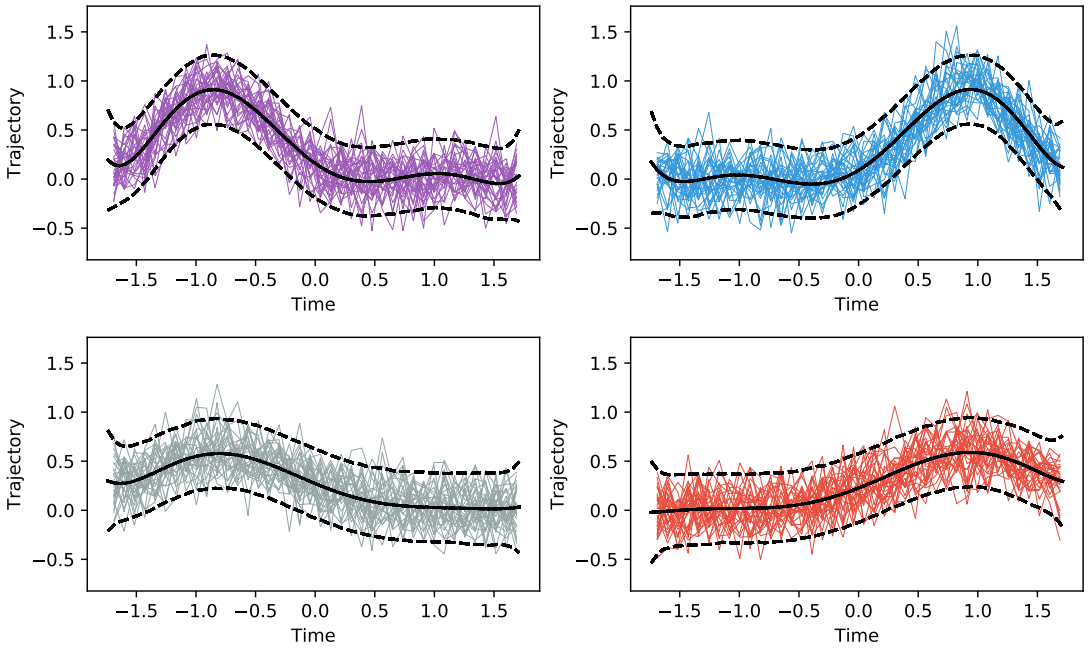


Figure 3: Data clustered via Bayes algorithm. The solid dark lines are the means of the predictive distributions on time, with prior parameters estimated by Bayes (case $\gamma = 0.41$). The dashed lines delimit the 95% empirical confidence regions.

selection criterion (BIC or ICL) was retained. The final estimates of Q are reported in Figure 4. BIC and asymptotic ICL selected the very same value of $Q = 2$ for hddc. The green curve reports the value of Q selected by the BIC for flexmix. When adopting ICL, the same values of Q were selected except for the fifth value of γ ($\gamma = 0.51$) where ICL selected $Q = 3$ instead of $Q = 4$. So we only reported the best choice. In the light of Figure 2, it is surprising that hddc never detects more then $Q = 2$ clusters, since its clustering abilities are comparable to those of the other algorithms. On the contrary, flexmix tends to overestimate Q for smaller values of γ, both with BIC or ICL. We re-run flexmix several times with different initial values of Z, random or k-means, but, at some point it always selects a value of Q higher than 4. As expected, the exact ICL from Eqs (8)-(15)-(16) selects $Q = 4$ groups for small values of γ and no algorithm or model selection criterion selects more than two clusters since $\gamma = 0.6$ onward.

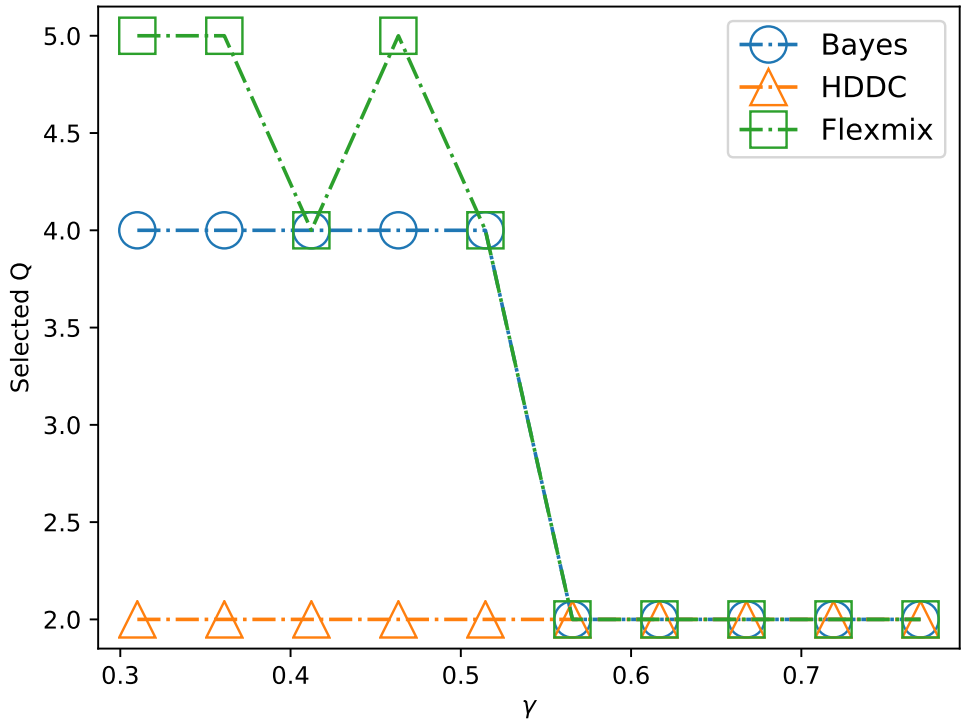


Figure 4: The estimated number of groups with decreasing γ.

5 Discussion and perspectives

The experiments in the previous section seem to suggest that, whereas our approach reaches the state of art in terms of clustering, it could bring a net improvement in terms of choice of Q. Indeed, practitioners might wish to use it only as a model selection tool, by simply using either the k-means Z or their own Z in Eq. (5) and (possibly) estimating other parameters via the MS alone. Further experiments on simulated and real datasets are needed to confirm such intuitions, especially in case of non-Gaussian noise.

Some final remarks. In the initialization of the prior parameters, when using a random initialization for Z, the choice of the Dirichet parameter α had an impact on the final clustering. For simplicity we assumed the Dirichlet distribution in Eq. 11 is symmetric. We noticed that the Classification Step detailed in Section 3.3 tended to switch too many observations at once, thus emptying some clusters. A similar remark was previously made by Côme and Latouche (2015); Corneli et al. (2016) in the context of graph clustering. However, such a tendency to

over-switching can be resisted by picking a large value for the initial α (see Ch.4 of Fruhwirth-Schnatter et al., 2019, for a detailed discussion about the Dirichlet prior hyper-parameters in mixture analysis.). In our our case, we adopted an initial value of $\alpha = 100$. The initial value of α is less relevant when the algorithm is provided with a "smarter" k-means initialization.

As anticipated in the previous section, the potential of the MS is currently limited due to a lack of regularity of the loss function. Further researches might focus on possible solutions to this issue. Moreover, we assumed that measurement times are shared across the observations $y_1, \dots, y_N$. Although the Bayesian strategy for clustering and model selection proposed in this paper does not need this assumption, the results presented in Section 3.3 do. A deep inspection of the general case, where the design matrix Φ in not the same for all the observations might be considered.

A ICL and BIC

We consider a standard mixture model where the N observed variables are denoted by $X := (x_1, \dots, x_N)$ and the corresponding latent variables by $Z = (z_1, \dots, z_N)$. The unknown number of mixing components is K. The (asymptotic) **BIC** and **ICL** criteria are defined as follows

$$BIC_K := \max_\theta \log p(X|\theta, K) - \frac{\nu(K)}{2} \log N \tag{24}$$

and

$$ICL_K := \max_\theta \log p(X, Z|\theta, K) - \frac{\nu(K)}{2} \log N, \tag{25}$$

where θ denotes the set of the model parameters, $\nu(K)$ is the number of model parameters and $\log p(\cdot)$ denotes the log density of the observations.

Remark 5. *The ICL criterion in Eq.* (25) *is an approximation of the marginal log-likelihood*

$$\log p(X, Z|K) = \log \int_\theta p(X, Z|\theta, K) p(\theta|K) d\theta, \tag{26}$$

where the model parameters are integrated out (Biernacki et al., 2000). If the prior distribution $p(\theta|K)$ is conjugated, the quantity on the left hand side of the above equation can be computed explicitly. We sometimes call it ***exact ICL****.*

The following notations are adopted

$$\hat{\theta} = \arg\max_\theta \log p(X|\theta, K), \tag{27}$$

$$\overline{\theta} = \arg\max_\theta \log p(X, Z|\theta, K) \tag{28}$$

and we stress that, in general, $\hat{\theta} \neq \overline{\theta}$.

The following Proposition formally shows that ICL_K in Eq. (25) is a lower bound of BIC_K. This result was mentioned in (Biernacki et al., 2000; Baudry et al., 2010) but not formally proven.

Proposition 2.

$$ICL_K \leq BIC_K.$$

Proof.

$$\begin{aligned} ICL_K - BIC_K &= \log p(X, Z|\overline{\theta}) - \log p(X|\hat{\theta}) \\ &= \log \frac{p(X, Z|\overline{\theta})}{p(X|\hat{\theta})} \\ &= \log \frac{p(X, Z|\overline{\theta}) p(X|\hat{\theta})}{p(X|\overline{\theta}) p(X|\hat{\theta})} \\ &\log p(Z|X, \overline{\theta}) + \log \frac{p(X|\overline{\theta})}{p(X|\hat{\theta})} \leq 0, \end{aligned}$$

where the dependence on K was omitted for simplicity and the last inequality comes from the discrete nature of the random variables Z_i and the definition of $\hat{\theta}$ and $\overline{\theta}$. □

B Linear algebra results

Theorem 2 (Silvester (2000))**.** *Let R be a commutative subring of $F^{n\times n}$, where F is a field and $F^{n\times n}$ denotes the set of matrices $n \times n$ over F. Let $M \in R^{m\times m}$. Then*

$$\det_F M = \det_F \left(\det_R(M) \right).$$

Lemma 1. *Consider a $\mathbb{R}^{N\times N}$ square matrix A such that*

$$A_{ij} = \begin{cases} a + \epsilon & \text{if} \quad i = j \\ a & \text{otherwise} \end{cases}$$

where a, ϵ are two real constants. Then

$$\det(A) = \epsilon^{N-1}(\epsilon + Na). \tag{29}$$

Proof. We proceed by recurrence. For $N = 1$, $A = (a + \epsilon)$ and Eq. (29) is verified. Now, let us assume that Eq. (29) holds for all $i \leq N$. The case where N is an even number is considered at first. Thus

$$\det M_{N+1} = \det \underbrace{\begin{pmatrix} a+\epsilon & a & \dots & a \\ a & a+\epsilon & \dots & a \\ \vdots & & \ddots & \vdots \\ a & a & \dots & a+\epsilon \end{pmatrix}}_{\text{N +1 columns}}$$

$$= (a+\epsilon)\det(M_N) - aN \det \underbrace{\begin{pmatrix} a & a & \dots & a \\ a & a+\epsilon & \dots & a \\ \vdots & & \ddots & \vdots \\ a & a & \dots & a+\epsilon \end{pmatrix}}_{\text{N columns}}$$

$$= (a+\epsilon)\det(M_N) - a^2 N \det(M_{N-1}) + a^2 N(N-1) \det \underbrace{\begin{pmatrix} a & a & \dots & a \\ a & a+\epsilon & \dots & a \\ \vdots & & \ddots & \vdots \\ a & a & \dots & a+\epsilon \end{pmatrix}}_{\text{N - 1 columns}}$$

and pursuing the recursion we obtain

$$\begin{aligned} \det M_{N+1} &= (a+\epsilon)\det(M_N) \\ &\quad - a^2 \left(\sum_{i=0}^{N-2} (-a)^i \frac{N!}{(N-i-1)!} \det(M_{N-i-1}) \right) \\ &\quad - a^{N+1} N!, \end{aligned} \tag{30}$$

where the sign of the last term on the right hand side (r.h.s.) of the equality is due to the assumption of an even N. Thanks to the inductive assumption, the first term on the r.h.s of the equality is

$$\begin{aligned} (a+\epsilon)\det(M_N) &= (a+\epsilon)\epsilon^{N-1}(\epsilon + Na) \\ &= a\epsilon^{N-1}(\epsilon + Na) + \epsilon^N(\epsilon + Na) \\ &= \epsilon^N(\epsilon + (N+1)a) + a^2 N \epsilon^{N-1}. \end{aligned}$$

Similarly, the inductive assumption can bu used to replace $\det(M_{N-i-1}) = \epsilon^{N-i-2}(\epsilon - (N-i-1)a)$ in the second term on the r.h.s. of Eq. (30). Thus, by developing

the sum over i we obtain

$$\begin{aligned}\det(M_{N-1}) &= \epsilon^N(\epsilon + (N+1)a) \\ &+ \cancel{a^2 N\epsilon^{N-1}} - \cancel{a^2 N\epsilon^{N-1}} \\ &+ \cdots + \cancel{a^N \epsilon N!} - \cancel{a^N \epsilon N!} \\ &= \epsilon^N(\epsilon + (N+1)a).\end{aligned}$$

The case where N is odd is analogous and the lemma is proven. □

Lemma 2. *Consider a $\mathbb{R}^{N\times N}$ square matrix A such that*

$$A_{ij} = \begin{cases} a+\epsilon & \text{if} \quad i=j \\ a & \text{otherwise} \end{cases}$$

where a, ϵ are two real constants. Then the inverse A^{-1} is

$$A^{-1}_{ij} = \begin{cases} \frac{\epsilon+(N-1)a}{\epsilon(\epsilon+Na)} & \text{if} \quad i=j \\ -\frac{a}{\epsilon(\epsilon+Na)} & \text{otherwise} \end{cases} \tag{31}$$

Proof. It suffices to verify that, given A^{-1} in Eq. (31), $AA^{-1} = I_N$. □

Theorem 3. *Consider an invertible square block matrix $M \in \mathbb{R}^{DN\times DN}$ such that*

$$M = \left(\begin{array}{c|c|c|c} B & A & \dots & A \\ \hline A & B & \dots & A \\ \hline A & A & \ddots & A \\ \hline A & \dots & \dots & B \end{array}\right) \tag{32}$$

where the non diagonal blocks $A \in \mathbb{R}^{D\times D}$ are symmetric matrices and the diagonal blocks are $B = A + I_D$. Then, the inverse matrix M^{-1} is still a block matrix

$$M^{-1} = \left(\begin{array}{c|c|c|c} V & W & \dots & W \\ \hline W & V & \dots & W \\ \hline W & W & \ddots & W \\ \hline W & \dots & \dots & V \end{array}\right)$$

where the non-diagonal blocks W are

$$W = -(I_D + NA)^{-1}A, \tag{33}$$

and the diagonal blocks V are

$$V = (I_D + NA)^{-1}(I_D + (N-1)A) = W + I_D. \tag{34}$$

Proof. We need to prove that MM^{-1} is a block matrix whose non-diagonal blocks are matrices in $\mathbb{R}^{D\times D}$ having zero everywhere (henceforth denoted by $\mathbf{0}_{\mathbb{R}^{D\times D}}$) and whose diagonal blocks are I_D. We observe that $(I_D+NA)^{-1} = \left(\overline{\det}(M)\right)^{-1}$, where $\overline{\det}(\cdot)$ is defined in Eq. (19) and $(I_D + NA)$ is invertible thanks to the assumption of invertible M combined with Lemma 1 and Theorem 1 in Silvester (2000). Now, the following notation is introduced

$$D_M := (I_D + NA)^{-1}$$

to simplify the exposition. Moreover, with a slight abuse of notation we denote by $(MM^{-1})_{ij}$ the block (and not the real entry!) at position (i, j) in MM^{-1}. Then, for $j \neq i$

$$\begin{aligned}(MM^{-1})_{ij} &= -BD_MA + AD_M(I_D + (N-1)A)\\ &\quad - (N-2)AD_MA\\ &= -D_MA + AD_M\\ &= \mathbf{0}_{\mathbb{R}^{D\times D}},\end{aligned}$$

where the last equality comes from

$$\begin{aligned}A(I_D + NA) = (I_D + NA)A \qquad &\Rightarrow\\ A = (I_D + NA)A(I_D + NA)^{-1} \;&\Rightarrow\\ (I_D + NA)^{-1}A = A(I_D + NA)^{-1}.&\end{aligned} \tag{35}$$

Similarly, for $i = j$

$$(MM^{-1})_{ii} = BD_M(I_d + (N-1)A) - (N-1)AD_MA$$

and some easy calculations show that the above quantity is equal to I_D. □

References

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., Gottardo, R., 2010. Combining mixture components for clustering. Journal of computational and graphical statistics 19 (2), 332–353.

Bergé, L., Bouveyron, C., Girard, S., et al., 2012. Hdclassif: An r package for model-based clustering and discriminant analysis of high-dimensional data. Journal of Statistical Software 46 (i06).

Bergé, L. R., Bouveyron, C., Corneli, M., Latouche, P., 2019. The latent topic block model for the co-clustering of textual interaction data. Computational Statistics & Data Analysis 137, 247 – 270.
URL http://www.sciencedirect.com/science/article/pii/S0167947319300726

Biernacki, C., Celeux, G., Govaert, G., 2000. Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans. Pattern Anal. Machine Intel 7, 719–725.

Birgé, L., Massart, P., 2007. Minimal penalties for gaussian model selection. Probability theory and related fields 138 (1-2), 33–73.

Bouveyron, C., Girard, S., Schmid, C., 2007. High-dimensional data clustering. Computational Statistics & Data Analysis 52 (1), 502–519.

Bouveyron, C., Latouche, P., Zreik, R., 2016. The stochastic topic block model for the clustering of vertices in networks with textual edges. Statistics and Computing.
URL https://hal.archives-ouvertes.fr/hal-01299161

Celeux, G., Govaert, G., 1991. A classification em algorithm for clustering and two stochastic versions. Computational Statistics Quaterly 2 (1), 73–82.

Côme, E., Latouche, P., 2015. Model selection and clustering in stochastic block models based on the exact integrated complete data likelihood. Statistical Modelling 15 (6), 564–589.

Corneli, M., Latouche, P., Rossi, F., 2016. Exact icl maximization in a non-stationary temporal extension of the stochastic block model for dynamic networks. Neurocomputing 192, 81–91.

Fruhwirth-Schnatter, S., Celeux, G., Robert, C. P., 2019. Handbook of mixture analysis. Chapman and Hall/CRC.

Muthén, B., Asparouhov, T., 2008. Growth mixture modeling: Analysis with non-gaussian random effects. Longitudinal data analysis 143165.

Muthén, B., Shedden, K., 1999. Finite mixture modeling with mixture outcomes using the em algorithm. Biometrics 55 (2), 463–469.

Rand, W. M., 1971. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association 66 (336), 846–850.

Silvester, J. R., 2000. Determinants of block matrices. The Mathematical Gazette 84 (501), 460–467.