Advances in Data Analysis and Classification (2023) 17:323–345
https://doi.org/10.1007/s11634-022-00501-x

REGULAR ARTICLE

Modal clustering of matrix-variate data

Federico Ferraccioli[1] · **Giovanna Menardi**[1]

Received: 25 March 2021 / Revised: 20 December 2021 / Accepted: 28 March 2022 /
Published online: 5 May 2022

Abstract
The nonparametric formulation of density-based clustering, known as modal clustering, draws a correspondence between groups and the attraction domains of the modes of the density function underlying the data. Its probabilistic foundation allows for a natural, yet not trivial, generalization of the approach to the matrix-valued setting, increasingly widespread, for example, in longitudinal and multivariate spatio-temporal studies. In this work we introduce nonparametric estimators of matrix-variate distributions based on kernel methods, and analyze their asymptotic properties. Additionally, we propose a generalization of the mean-shift procedure for the identification of the modes of the estimated density. Given the intrinsic high dimensionality of matrix-variate data, we discuss some locally adaptive solutions to handle the problem. We test the procedure via extensive simulations, also with respect to some competitors, and illustrate its performance through two high-dimensional real data applications.

Keywords Matrix-variate data · Modal clustering · Mean-shift · Kernel density · Nearest neighbors

Mathematics Subject Classification Primary 62G05 · Secondary 62H30

1 Introduction

The analysis of complex data in the form of matrices represents an active area of research. Classical examples are longitudinal studies and multivariate spatio-temporal data, where statistical observations are represented by vectors of variables, measured on subjects over different times or locations, or modern analyses where the matrix structure is intrinsic to the problem, as in the case of image data, adjacency matrices

✉ Federico Ferraccioli
federico.ferraccioli@unipd.it

Giovanna Menardi
menardi@stat.unipd.it

[1] Dipartimento di Scienze Statistiche, Università degli Studi di Padova, Padua, Italy

representing networks, covariance or similarity structures etc. While the focus is usually intended to the former case, collections of data of such type are often referred to as *three-way*, with two ways associated to the row and column dimension of each matrix-variate observation and the third one represented by subjects. For an extensive review, see Kroonenberg (2008).

A vast body of literature has focused on the development of supervised methods for matrix-valued data. Some accounts with specific focus on this topic are Viroli (2012) and Ding and Cook (2018), yet under the same umbrella we shall include regression models with multivariate observations gathered over time (see, e.g. Diggle et al. 2002, for a standard account) or penalized linear regression models with matrix-valued response (Zhou and Li 2014). Specific classes of matrix-valued data, such as semi-definite positive matrices (see for example Dryden et al. 2009) and orthogonal matrices (see for example Chakraborty and Vemuri 2019), have also received recent attention. Far less attention has been devoted to the unsupervised case. Historical contributions here are the Tucker3 and the Candecomp/Parafac models (Kroonenberg 2008, Ch 5.7 and, respectively, 4.6) and developments, as well as various attempts to reduce data dimension via principal component analysis or akin methods (see, e.g., Sakata 2016, Ch. 1 and 3).

For the aim of clustering, longitudinal or functional data methods have been largely proposed for grouping subjects. Among these, it is worth to mention the ones accounting for the observation of multiple features measured over time (see, e.g., Schmutz et al. 2020; Jacques and Preda 2014), yet only a few contributions refer to the more general case of three-way structures, where the possible dependency along the column dimension does not necessarily derive from a specific order of the observations. Stemming from Basford and McLachlan (1985), Viroli (2011) develops model-based clustering building on mixtures of matrix-variate Normal distributions and its bayesian counterpart (Viroli 2011). Similarly, Gallaugher and McNicholas (2018) propose skewed matrix-variate distributions for unsupervised and semi-supervised classification. More recent proposals are Tomarchio et al. (2020), that introduce two matrix-variate distributions for model-based clustering, and Sarkar et al. (2020), that develop parsimonious mixture models.

With a somewhat different aim in mind, further scattered examples which are worth to mention are Wang et al. (2019), which perform tensor decomposition to cluster individuals and tissues in gene expression data; Vermunt (2007), where a hierarchical mixture model is used to group longitudinal matrix-variate data in clusters which possibly vary across times, and Vichi et al. (2007) who perform clustering of subjects and factorial dimensionality reduction of variables and occasions of three-way data.

In this work we extend the *nonparametric* formulation of density-based clustering, also known as *modal* clustering, to the framework of matrix-variate data. Here, clusters are identified as the "domains of attraction" of the modes of the true density underlying the data (Stuetzle 2003). The inherent notion of cluster is hence not linked to any predefined shape, and determining the number of clusters is an integral part of the estimation procedure. While with a different rationale, nonparametric clustering shares with its more widespread parametric counterpart a sound probabilistic foundation, which also allows for a precisely defined population goal. In fact, the issue of density estimation, usually addressed via nonparametric methods, assumes a key role

in order to approximate the ideal population goal of modal clustering, along with the operational search of the modal regions.

After providing an overview on the modal clustering approach, we introduce a kernel estimator for matrix-variate density functions. We then study its asymptotic properties, also with reference to the problem of optimal bandwidth selection. Due to the intrinsic high dimensionality of matrix-variate data, which impacts on both the accuracy of the estimate and the computational complexity, we explore some local solutions to handle the problem. Additionally, we propose an extension of the mean-shift procedure for the identification of the modes of the estimated density. Finally we perform an extensive simulation study and illustrate the performance of the proposed method on two sets of real data.

2 An overview on modal clustering

In the following, and throughout the paper, we will denote by lower-case symbols both scalar and vector-valued objects, whereas matrices will be denoted by uppercase letters. With some abuse, we will also use the same notation to indicate random objects and associated realizations, and specify explicitly their nature when it is not clear from the context. In a standard multivariate setting, *modal* clustering relies on the assumption that the observed data $\mathcal{X} = (x_1, \ldots, x_N)$ are realizations of a multidimensional random variable $x \in \mathbb{R}^P$ with (unknown) probability density function f. The modes of f are regarded to as representatives of the clusters, which are in turn represented by their domains of attraction. Broadly speaking, if the underlying density is gured as a mountainous landscape, and modes are its peaks, clusters are the 'regions that would be ooded by a fountain emanating from a peak of the mountain range' (Chacón 2015). Morse theory allows a more formal framing of the problem, by defining clusters as the stable manifolds of the gradient flow associated with the local maxima of f. These are represented by the sets of all the points which converge to the same mode by following the gradient ascent paths of the true density.

While the population clustering goal is defined precisely in terms of features of the underlying density, this is in practice unknown, and needs to be estimated. The issue is far from being trivial, as the estimated density determines the modal regions, and hence governs the final clustering. A standard choice, within the class of nonparametric methods, is the kernel density estimator

$$\hat{f}(x; h) = \frac{1}{Nh^P} \sum_{n=1}^{N} K\left(h^{-1}(x - x_n)\right), \tag{1}$$

where the kernel K is a probability density on $\mathbb{R}^P$, symmetric around zero, and the bandwidth $h > 0$ is a scale parameter defining the degree of smoothing. While the choice of the kernel is known not to have a strong impact on the performance of $\hat{f}$, a proper selection of the bandwidth turns out to be crucial. Small values of h lead to an undersmoothed density estimate, with the possible appearance of spurious modes, whereas too large values result in an oversmoothed density estimate, possibly hiding relevant features.

A further aspect to account for in modal clustering is to operationally characterize the clusters as the domains of attraction of the density modes. Most of the contributions in this direction take their steps from the *mean-shift* algorithm (Fukunaga and Hostetler 1975) which, starting from a generic point $y^{(0)}$, recursively shifts it uphill to a local weighted mean of the data, along the direction of the gradient of its kernel estimate:

$$y^{(s+1)} = \sum_{n=1}^{N} w_{n,h}(y^{(s)})x_n.$$

The weights $w_{n,h}(\cdot)$ are specified as normalized components of the gradient of the kernel function. Hence, the mean shift is shown to be a gradient ascent algorithm based on a normalized kernel estimator of the gradient. The convergence of the sequence $\{y_0, y_1, \ldots, y_s, \ldots\}$ to a local mode of (1) has been studied under various assumptions by Ghassabeh (2015) and Arias-Castro et al. (2016).

A partition of the data is finally obtained by grouping in the same cluster the observations ascending to the same mode of the density. The reader may refer to Menardi (2016) and references therein for insights on modal clustering.

3 Matrix-variate extension of modal clustering

3.1 Kernel density estimation of matrix-variate data

Let $X_1, \ldots, X_N$ be a sample of *i.i.d.* realizations of a $P \times T$ random matrix $X = \{x_{p,t}\}_{p=1,\ldots,P,t=1,\ldots,T}$, which we shall assume to be defined on the vector space $M_{P\times T} \subseteq \mathbb{R}^{P,T}$. The (unknown) distribution of X is naturally described by some probability density function $f : M_{P\times T} \mapsto \mathbb{R}_+$, with $\int_{M_{P\times T}} f(X)\mathrm{d}X = 1$. Since we are dealing with a vector space, the integral is intended as the component-wise integral of f on its support.

Consider an integrable kernel $K : \mathbb{R}^{P,T} \mapsto \mathbb{R}_+$, with unit integral and spherically symmetric, i.e. $\int_{M_{P\times T}} XK(X)dX = 0$. We define the *kernel density estimator* for matrix-variate data as

$$\hat{f}(X; h) = \frac{1}{Nh^{P\cdot T}} \sum_{n=1}^{N} K(h^{-1}(X - X_n)), \qquad h > 0. \tag{2}$$

With the above established convention on defining matrix-variate integrals as their component-wise counterpart, and the same for derivatives, most of standard results on kernel density estimators extend naturally to the matrix-variate setting. The Mean Integrated Square Error (MISE) admits forthwith the usual representation (e.g. Chacón and Duong 2018, p. 28)

$$\begin{aligned}\mathrm{MISE}(\hat{f}(X; h)) &= \mathbb{E}\int_{\mathbb{R}^{P,T}} (\hat{f}(X; h) - f(X))^2 dX \\ &= \int_{\mathbb{R}^{P,T}} \mathrm{Var}(\hat{f}(X; h))dX + \int_{\mathbb{R}^{P,T}} \mathrm{Bias}^2(\hat{f}(X; h))dX\end{aligned}$$

$$= \mathrm{IV}(\hat{f}(X;h)) + \mathrm{ISB}(\hat{f}(X;h)) \tag{3}$$

where $\mathrm{IV}(\cdot)$ is the integrated variance and $\mathrm{ISB}(\cdot)$ is the integrated squared bias of the estimator. Its dependence on the bandwidth is nonetheless not easily disclosed, as the latter enters implicitly via the integrals involving the kernel. To highlight the effect of the bandwidth, it is useful to derive an asymptotic approximation of the MISE. To this aim, we further assume the following:

(i) f is square integrable and twice differentiable, with all its second order partial derivatives bounded, continuous and square integrable;
(ii) The kernel K is, in turn, square integrable, with finite second order moments $\int X \otimes XK(X)dX = m_2(K)\mathrm{vec}\mathbb{I}_{P\times T}$, and $m_2(K) = \int x_{p,t}^2 K(X)dX$, $p = 1,\ldots,P, t = 1,\ldots T$. The symbol $\otimes$ here denotes the Kronecker product, while $\mathbb{I}_d$ denotes the d-dimensional identity matrix;
(iii) The bandwidths $h = h_N$ form a positive sequence, such that $h \to 0$ and $N^{-1}h \to 0$ as $N \to \infty$.

Then, the following holds.

Proposition 1 *The asymptotic mean integrated squared error (AMISE) for $\hat{f}(\cdot;h)$ is*

$$AMISE(\hat{f}(\cdot;h)) = N^{-1}h^{-(P\cdot T)}R(K) + \frac{1}{4}h^4 m_2(K)^2 R(\Delta f),$$

and it is minimized by

$$h_{AMISE} = \left(\frac{(P\cdot T)R(K)}{m_2(K)^2 R(\Delta f)}\right)^{\frac{1}{(P\cdot T)+4}} N^{-\frac{1}{(P\cdot T)+4}}, \tag{4}$$

where, for a square integrable function $a : \mathbb{R}^{P\times T} \mapsto \mathbb{R}$ we denote by $\Delta a = \sum_{p=1}^{P}\sum_{t=1}^{T}\frac{\partial^2 a(X)}{\partial x_{p,t}^2}$ the Laplacian operator and $R(a) = \int_{\mathbb{R}^{P,T}} a(X)^2 dX$ its square integral.

Proof See "Appendix". □

Hence, as for vector-valued data, the approximately optimal bandwidth converges to zero as N increases at the rate $N^{-\frac{1}{(P\cdot T)+4}}$. Also, the optimal solution (4) relies on the knowledge of the true f, and hence cannot be directly used to define the optimal smoothing amount. Consistently with the vector case, automatic bandwidth selection can be built by first estimating either the MISE or its asymptotic version (AMISE), and then minimising such estimate to yield a bandwidth computed solely from the data. Standard approaches based on cross-validation, bootstrap, or based on replacing the target density with a given parametric model in the expressions of the MISE/AMISE can be easily extended to the matrix-variate framework.

In fact, as for the standard multivariate settings, the use of a scalar bandwidth h may result in a poor flexibility, and richer classes of parameterizations may be alternatively considered. The maximal extent of flexibility would require the awkward use of a

four-way structure whose entries would reflect all the possible covariances between pairs of the X components. Alternatively, the vectorization operator may be easier to this aim, by mapping $\mathbb{R}^{P,T}$ to $\mathbb{R}^{P\cdot T}$ and stacking the column vectors of X underneath each other in order from left to right. With this representation, a full, unconstrained bandwidth H takes the form of a symmetric, semidefinite matrix $P \cdot T \times P \cdot T$, yet with some limitations from the algebraic and computational points of view.

A remarkable simplification may be induced by certain Kernels, for which a separable structure of H is available, so that an equivalent specification represents the matrix-variate $P \times T$ Kernel as a special case of a $PT-$variate Kernel with bandwidth $H = U \otimes V$, with U and V symmetric positive definite matrices of dimension $P \times P$ and $T \times T$, respectively. Elliptical models belong to this family, and are defined as

$$K(V^{-1/2}(X - X_n)U^{-1/2}) = |V|^{-\frac{P}{2}}|U|^{-\frac{T}{2}} g\left(-\frac{1}{2}\text{tr}(V^{-1}(X - X_n)^\top U^{-1}(X - X_n))\right),$$

where $g : \mathbb{R} \mapsto \mathbb{R}_+$ is such that $\int_{\mathbb{R}_+} z^{pt-1} g(z^2)dz < \infty$ (Caro-Lopera et al. 2016). The matrices U and V act independently on the rows and columns and are easier to handle than a full specification of the matrix H, that might result challenging even when P and T are small. By following the same steps as in Proposition 1, we may obtain the expression of the AMISE when $H = U \otimes V$. In this case, the first term depends on the determinant $|H|^{-1/2}$ instead of $h^{-(P\times T)}$, while the second term involves the full Hessian instead of the simple Laplacian. The general result, however, does not lead to an explicit formula for the optimal bandwidth matrix.

Note that the simplest case, where $U = h_U \mathbb{I}_P$ and $V = h_V \mathbb{I}_T$, reduces to the form

$$K((h_U h_V)^{-1}(X - X_n)) = h_U^{-P} h_V^{-T} g\left(-\frac{1}{2(h_U h_V)^2}\text{tr}((X - X_n)^\top (X - X_n))\right),$$

hence the choice of two distinct smoothing parameters for rows and columns brings back to the scalar case as an effect of the separability of the scale matrix H.

Within the class of elliptical kernels, we may set $g(\cdot) = (2\pi)^{-\frac{P\cdot T}{2}} \exp(\cdot)$ and obtain the matrix Normal density, a natural candidate for the kernel function which plays a pivotal role in the matrix-variate framework, as for the univariate and multivariate settings (see, e.g, Gupta and Nagar 2018).

3.1.1 Adaptive kernel

As an overall problem shared by nonparametric tools, kernel estimators are known to strongly suffer from the curse of dimensionality. On one side, the required sample size to achieve an acceptable accuracy becomes disproportionately large as the dimensions increases, leading to intractable problems, even computationally. On the other side, in high dimensions, the sparsity of data leads much of the probability mass to flow to the tails of the density, possibly averaging away features in the highest density regions and giving rise to the birth of spurious modes.

These arguments could discourage from the application of modal clustering on matrix-variate data, which are intrinsically high-dimensional, except for very small

values of P and T. In fact, nonparametric estimates can still be useful to coarsely describe the data structure, and often allowing different amounts of smoothing is advisable to capture local structures of the data. In this direction, adaptive estimators build on the idea that for data-sparse regions, a large bandwidth is needed to compensate for the few nearby data points, and conversely, for data-dense regions, a small bandwidth applies smoothing in a small region due to the presence of many nearby data points. As a general principle, we may distinguish between *balloon* and *sample point* estimators, which replace h in equation (2) by $h(X)$ and $h(X_i)$ respectively. See Chacón and Duong (2018, §4.1) for an overview in the multivariate setting. Within these classes, we consider, for the matrix-variate setting, a k-nearest neighbor (k-NN) extension of the two estimators, defined as

$$\hat{f}_B(X;k) = \frac{1}{N\delta_k(X)^{P\cdot T}} \sum_{n=1}^{N} K\left(\delta_k(X)^{-1}(X - X_n)\right), \tag{5}$$

$$\hat{f}_{SP}(X;k) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{\delta_k(X_n)^{P\cdot T}} K\left(\delta_k(X_n)^{-1}(X - X_n)\right), \tag{6}$$

where $\delta_k(X) = ||X - X^{(k)}||_F$ is the Frobenius distance of X from its k-th nearest neighbour $X^{(k)}$.

3.2 Mean shift for matrix-variate data

Once the density has been estimated via the matrix-variate extension discussed so far, clusters can be associated to the domains of attraction of the modes of such density, to be intended as high-density subsets of the sample space surrounding the (matrix-variate) local maxima of the density.

With this regard, the following proposition states that the hill-climbing property of the mean-shift algorithm still holds in the matrix-variate setting.

Proposition 2 *Consider a differentiable kernel $K : \mathbb{R}^{P,T} \mapsto \mathbb{R}_+$, with unit integral, and spherically symmetric. Let $\kappa(\cdot) : \mathbb{R}_+ \mapsto \mathbb{R}$ be a function such that $K(X) = \frac{1}{2}\kappa(tr(X^\top X))$ and its derivative $\kappa'(u) \leq 0$.*

Then, starting at $Y^0 \in \mathbb{R}^{P,T}$, the sequence defined by

$$Y^{(s+1)} = Y^{(s)} + M(Y^{(s)}) = \sum_{n=1}^{n} w_{n,h}(Y^{(s)}) X_n \tag{7}$$

describes a gradient ascent algorithm on (2), with

$$M(Y) = \frac{\sum_{n=1}^{N} \kappa'\left(h^{-2}\mathrm{tr}((X_n - Y)^\top (X_n - Y))\right)(X_n - Y)}{\sum_{n=1}^{n} \kappa'\left(h^{-2}\mathrm{tr}((X_n - Y)^\top (X_n - Y))\right)} \tag{8}$$

denoting the mean-shift and

$$w_{n,h}(Y) = \frac{\kappa'\left(h^{-2}\mathrm{tr}((X_n - Y)^\top (X_n - Y))\right)}{\sum_{n=1}^{N} \kappa'\left(h^{-2}\mathrm{tr}((X_n - Y)^\top (X_n - Y))\right)}.$$

Proof See "Appendix". □

While loosing its interpretation as an iterative weighted average of the observations, the gradient ascent nature of the mean-shift may be derived also for more complex structures of the bandwidths. When a kernel function with separable structure $H = U \otimes V$ is used, for instance, the (8) becomes

$$M(Y) = \frac{\sum \kappa'\left(\mathrm{tr}(V^{-1}(X_n - Y)^\top U^{-1}(X_n - Y))\right) V^{-1/2}(X_n - Y)U^{-1/2}}{\sum \kappa'\left(\mathrm{tr}(V^{-1}(X_n - Y)^\top U^{-1}(X_n - Y))\right)}$$

with some simple mathematical manipulation.

With respect to the adaptive estimator (6), the same proposition holds, with the only caution of replacing h with $\delta_k(X_n)$. Conversely, the same arguments do not generally apply to the balloon estimator (5), since the kernel depends on X also through $\delta_k(X)$, and therefore does not allow to derive a general expression for its gradient. An exception in the multivariate case occurs when the kernel is chosen among the beta family, where the problem simplifies remarkably (Duong et al. 2016). The same naturally extends to the matrix variate case. Specifically, when a uniform kernel on the unit PT-ball is selected, the gradient ascent property of the mean-shift is shown to hold with extreme computational efficiency, as stated by the following result.

Corollary 1 *Consider the adaptive estimator in* (5)*, with* $K(X) = \nu_0^{-1}\mathbb{1}\{X \in B_{PT}(0,1)\}$ *and* $B_{PT}(0,1)$ *the unit ball centered at* 0*, with hypervolume* ν_0*. Then, the mean shift sequence* (7) *takes the form*

$$Y^{(s+1)} = \frac{1}{k} \sum_{X_n : X_n \in B_{PT}(Y^{(s)}, \delta_k(Y^{(s)}))} X_n.$$

The proof follows the same steps of the one of Proposition 2. See also Duong et al. (2016) for the multivariate case.

4 Simulations

4.1 Settings

In this Section we present an extensive simulation study with the aim of evaluating the performances of the proposed approach to cluster three-way data, with respect to

the following aspects: (1) different group configurations, sample sizes, data dimension; (2) the use of different formulations of kernel-type matrix-variate estimators; (3) comparison with some competitors.

In the case of matrix-valued data, generating random samples with some interesting, nontrivial structure to be disclosed is awkward, and literature is quite scarce. Some Gaussian matrix-variate examples can be found in Viroli (2011) and Viroli (2011). Here we follow a different route, based on multidimensional Discrete Cosine Transform (DCT, Strang 1999), a transformation technique for data compression, widely used in digital media and imaging. DCT is able, in principle, to handle and control for structures with varying degrees of complexity. For each cluster, we define a matrix prototype M of size $P \times T$, and express it as

$$M = L^\top \Omega R,$$

where L and R are two orthogonal matrices with dimensions $P \times P$ and $T \times T$, respectively, that contain the basis of the decomposition. The matrix Ω, of dimension $P \times T$ is the so called DCT, and its elements $\omega_{p,t}$ are computed stemming from the entries $m_{p,t}$ of M as (see Makhoul 1980)

$$\omega_{p,t} = 4 \sum_{i=1}^{P} \sum_{j=1}^{T} m_{i,j} \cos\left(\frac{\pi(2i-1)(p-1)}{2P}\right) \cos\left(\frac{\pi(2j-1)(t-1)}{2T}\right).$$

The cosine factors are the elements of the matrices L and $R^\top$, respectively.

A random matrix X of size $P \times T$ is then built starting from M by replacing each of the entries $\omega_{p,t}$ of Ω by $\omega_{p,t} + \epsilon u$, with $\epsilon \sim N(0, \sigma^2)$ and $u \sim Bin(1, \rho)$. In practice, a random proportion ρ of DCT coefficients is contaminated with normal error of zero mean and fixed variance. The role of ρ is twofold: on one side, it determines, along with σ^2, the amount of sample variability and, on the other side, it governs the shape and distribution of the clusters. While setting ρ equal to one determines the generation of matrix-variate spherical normal clusters, any lower proportion leads to some departure from such distribution. Figure 1 presents a graphical example of matrix prototype, and three random realizations associated to increasing values of ρ.

Three main clustering configurations are considered: a single-group setting, defined by the prototype A illustrated in Fig. 2; a balanced two-groups setting, with matrix-variate data equally sampled from prototypes B and C of Fig. 2, and an imbalanced two-groups setting, with data again sampled from prototypes B and C in the uneven proportion 0.1 and 0.9, respectively. For each of these settings, varying sample size, data dimensions, cluster variability and distribution are evaluated, by letting $N \in \{1000, 3000\}$, $(P \times T) \in \{5\} \times \{5, 20\}$, and $\rho \in \{0.1, 0.3, 1\}$. For each setting, 500 Monte Carlo samples have been generated.

Modal clustering is performed via the mean-shift algorithm discussed in Section 3.2, applied to three different formulations of kernel estimator. A fixed bandwidth estimator is evaluated, with Normal matrix-variate kernel and scalar bandwidth set as asymptotically optimal to estimate the first derivative of a Normal matrix-variate density. While this choice is unarguably sub-optimal, especially in the presence of

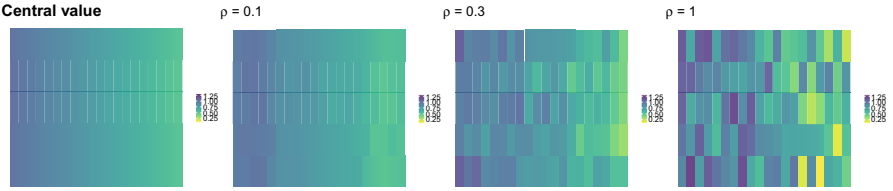

Fig. 1 Graphical representation of an example of matrix prototype M (left), where color intensities are associated to different values of the matrix entries. The second, third and fourth panel display three random realizations of the random matrix associated to M, for increasing values of ρ.

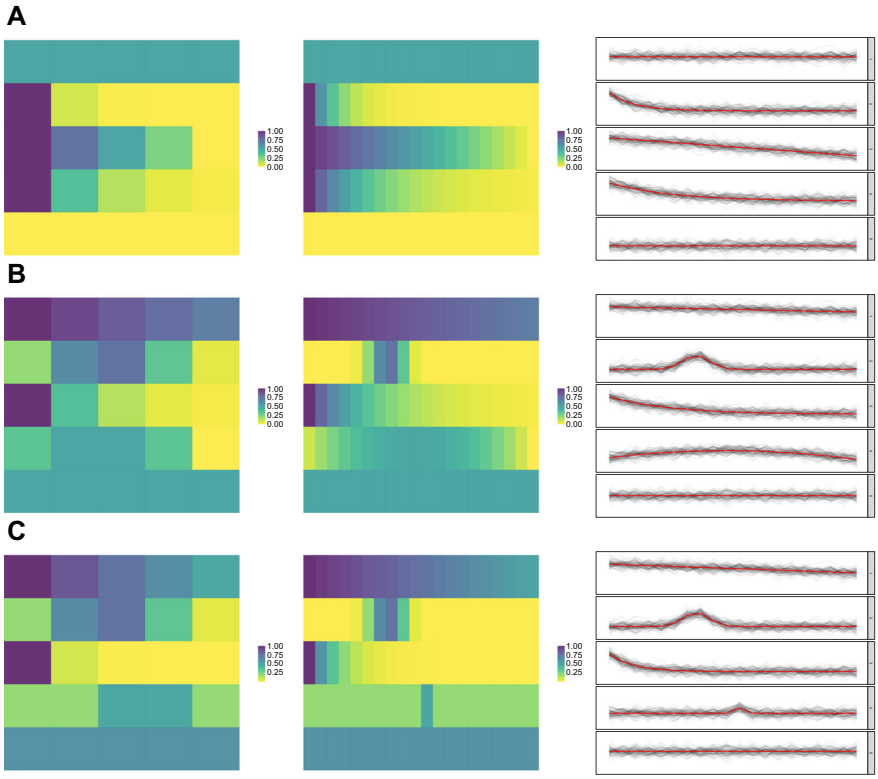

Fig. 2 Graphical representation of the three matrix prototypes used to define the groups in the various simulation settings. Cf. Fig. 2 for the left and middle panels, associated to the settings $(P \times T) = (5 \times 5)$ and $(P \times T) = (5 \times 20)$ respectively. The right panels display, for the $P = 5$ variables of each cluster prototype, a subsample of curves generated via DCT (black dashed lines) with $\rho = 1$ and over-imposed the associated prototype of each row (red solid line) (color figure online)

multimodal structures, it has been proven successful in many applications of modal clustering in the standard multivariate setting (Menardi 2016). In fact, since this rule of thumb is known to oversmooth the true density, it seems in principle a sensible choice in the presence of high dimensional data, where oversmoothing may relieve the problem of spurious cluster in the low density regions. As a representative of balloon

estimators, we consider the (5) with Uniform Kernel on the $PT-$ball of radius $\delta_k(\cdot)$, and $k \in (0.5\sqrt{N}, \sqrt{N}, 5\sqrt{N})$. Finally, we consider a sample point estimator (6), with Normal matrix-variate kernels, bandwidth $h\delta_k(X_n)$, $k \in (0.5\sqrt{N}, \sqrt{N}, 5\sqrt{N})$ and h set as in the fixed bandwidth case.

As a benchmark, we also perform clustering via K-means and model-based clustering based on mixtures of matrix-variate Normal distributions. In the former case the analysis is conducted via matrix vectorization, and the number of clusters is determined by using the best *Silhouette* score (Rousseeuw 1987) in the range of values $\{2, \ldots, 9\}$. In the latter case we implement the approach of Viroli (2011) and the number of mixture components is selected via the BIC within the range $\{1, \ldots, 9\}$.

To evaluate the quality of partitions with respect to a true labeling, the Adjusted Rand Index (ARI) is known to represent the mainstream criterion. However, when one of the compared partition is formed by one group only, the ARI returns its minimum value, disregarding the label distribution in the other partition. For this reason, we preferably consider the Fowlkes-Mallows index, which is sensitive to a different quality of partitions also when one of the two partitions is formed by one group only. Results deriving from the ARI evaluation are reported in the Supplementary Material for completeness. See, e.g. (Hennig et al. 2015, Ch. 27) for a review of the criteria to evaluate cluster quality.

All the analyses have been run in the `R` environment (R Core Team 2020), with the modal clustering routines built as suitable modifications of functions available in the `ks` packages (Duong 2019). The code is available at https://github.com/federicoferraccioli/kmsMatrix.

4.2 Results

Results referred to simulations of samples of size $N = 1000$ and the use of FM index are displayed in Figs. 3, 4 and 5. Modal clustering performs successfully in all the considered settings, yet with some not negligible differences. The balloon k-NN kernel estimator is the one which mostly offers guarantees of revealing the true modal structure, as in all the considered settings there exists at least one value of k, among the examined ones, leading to a very accurate cluster detection. Cluster quality is not much sensitive to the selected number of nearest neighbors, at least for low to moderate amount of variability. For large ρ, conversely, better results are achieved by a larger amount of smoothing, i.e. when k takes its largest value among the three examined ones. It is worth to note that such largest value of k produces an accurate classification of the observations in all the considered settings.

Estimating the data density with the use of a scalar bandwidth, as well as via the sample point estimator, results in a faithful cluster recovery in the presence of a small amount of variability. Increasing ρ, conversely, produces a progressive worsening of the results. A deeper insight of the results suggests that such lower accuracy is due to the arising of spurious clusters.

It is perhaps unexpected that increasing the matrix dimension not always reduces the quality of detected clusters. In fact, the larger sparsity of the data in higher dimensions increases cluster separation, *ceteris paribus*, hence the performance of modal

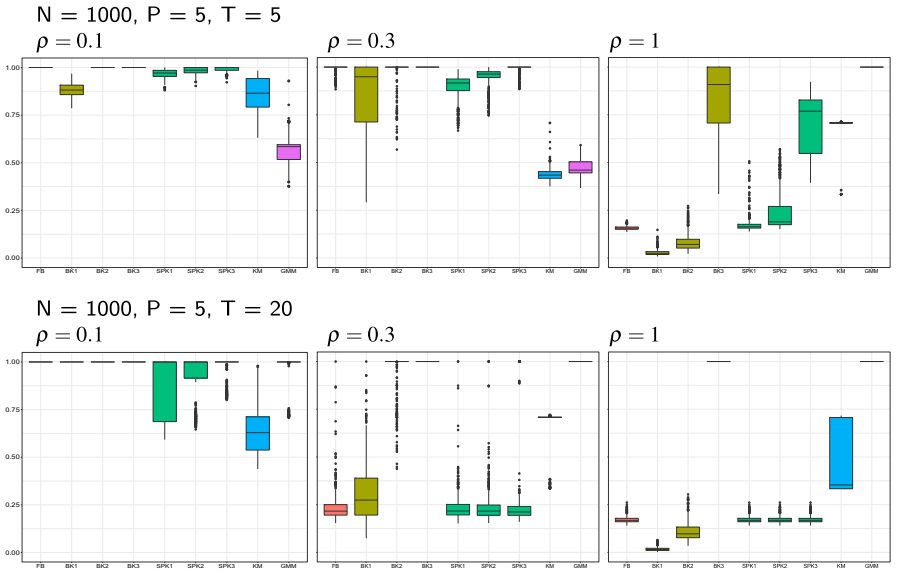

Fig. 3 Simulation results of the single group setting, with sample size, matrix dimension and proportion ρ of DCT coefficients defined in the figure headings. Each panel displays the Monte Carlo distribution of the FM index when modal clustering is run with a fixed bandwidth kernel estimator (FB), a balloon and a sample point k-NN estimator, both with increasing values of k (BK1, BK2, BK3, SPK1, SPK2, SPK3), and when K-means and model-based clustering with gaussian mixture models are run (KM and, respectively, GMM)

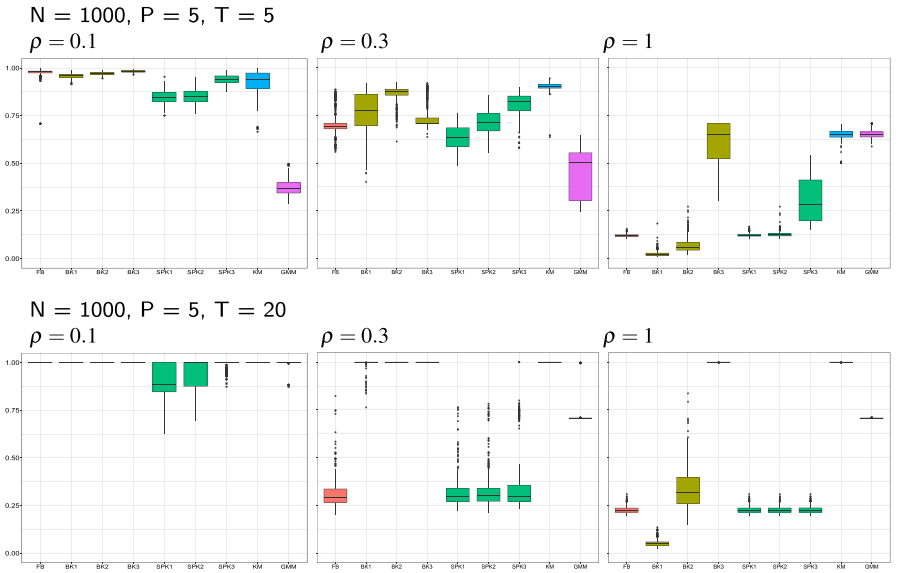

Fig. 4 Simulation results of the balanced groups settings. Cf Fig. 3

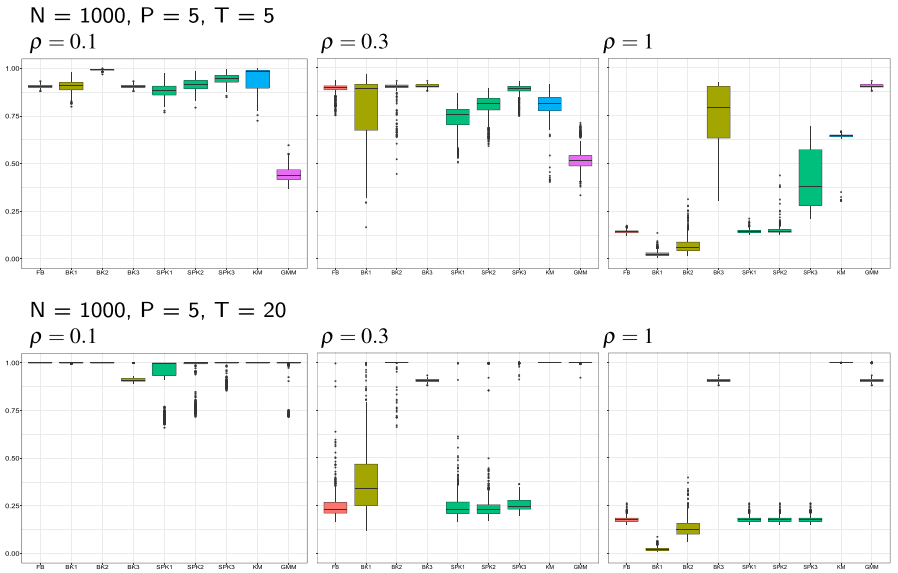

Fig. 5 Simulation results of the unbalanced groups settings. Cf Fig. 3

clustering tends to improve. For those situations where this is not true, usually associated to the use of a sample point estimator, the arising of spurious clusters is the main responsible for the worsened behavior, hence we guess that a larger amount of smoothing would relieve such behaviour. Note, in fact, that in the considered estimators based on the k-nearest neighbors, the examined rules of thumb to select k vary with the sample size and not with the data dimensionality. While further research should usefully shed light on specific criteria for bandwidth selection, we believe that relevant results have emerged from the analysis, confirming the opportunity of a satisfactory use of nonparametric density-based methods even in high dimensional spaces.

Despite its simplicity and its known limitations, K-means clustering produces overall notable results. In the two-groups settings the quality of the detected partitions ranges from fair to very good, and the clustering structure is roughly caught even when the amount of variability is high. The algorithm finds it harder in the unbalanced settings, yet it can anyway identify the gross clustering structure. Not surprisingly, the worst results refer to the single cluster settings. In fact, the number of clusters set in K-means is selected by maximizing the Silhouette index, which cannot, by construction, be evaluated when $K = 1$. Hence, all the K-means results refer to partitions formed by at least two clusters. The low values of the Fowlkes-Mallows index, however, suggest that whatever number of clusters is selected by the Silhouette score, in the single-group settings observations are allocated uniformly to the clusters, instead of favouring a single group. This result is, in fact, consistent with the usual behavior of K-means clustering which tends to split data into balanced groups. This reason, along with the lack of a formal criterion to determine the number of clusters, overall discourage from the use of K-means, similarly to the standard multivariate framework.

Model-based clustering is known to represent a generalization of K-means, where clusters are modeled to possibly vary in variance and proportion. Additionally, unlike

K-means, the use of the BIC allows for selecting single-cluster models. Despite these advantages, model-based clustering looks competitive in the $\rho = 1$ settings only, where clusters are designed to be Gaussian, consistently with the specified model. The performance of model-based clustering improve and get competitive when the data dimensionality increases, thus confirming the increase of cluster separation discussed above.

Results from using a larger sample size ($N = 3000$) are available in the Supplementary Material as they are essentially the same as the ones obtained with $N = 1000$. We believe that in high dimensional spaces as the ones here considered, to produce a remarkable improvement of the results, the sample size should increase to an unfeasible extent for simulation purposes.

Computational times are overall feasible in all the considered settings, yet with some non-negligible differences. Among the kernel based methods, the fixed-kernel and the sample-point variants are the most computationally expensive, especially with increasing sample sizes. This is due to the fact that at each step the kernel gradient is evaluated using all the observations. On the contrary, the baloon k-NN estimator shows excellent efficiency, as the evaluation is done only in a neighbourhood of each data point. The three procedures are far less affected by increasing matrix dimension. Although the k-means algorithm requires to be run for different values of K, its overall behaviour is comparable to the baloon k-NN estimator in terms of computational time. Unlike the other methods, the computational burden borne by Gaussian Mixture models is the most affected by increase of matrix dimension, which requires a larger number of parameters to be estimated. Similarly to k-means, Gaussian mixtures require to be estimated for different number of components before selecting the best model, which increases remarkably the computational times since the larger K, the larger number of parameters need to be estimated. Considerations above find numerical evidence in Table 1, reporting the average computational times in seconds, elapsed on a standard machine to run each of the competitors in the two-balanced clusters setting, for sample sizes and matrix dimensions considered in simulations.

5 Application

5.1 Activity Tracking

As first real data application, we consider a dataset describing a number of daily and sports activities measurements, detected by 5 sensors positioned on the torso, the wrists

Table 1 Average computational times in seconds, elapsed on a standard machine to run each of the competitors in the two-balanced clusters setting for sample sizes and matrix dimensions considered in simulations. Times to run modal clustering methods include evaluations of the amount of smoothing; k-means and GMM refer to times to produce partitions with 2 to 9 clusters

	FB	BK3	SPK3	KM	GMM
$N = 1000, P = 5, T = 5$	2.47	0.24	6.90	0.40	6.18
$N = 1000, P = 5, T = 20$	2.76	0.67	9.87	1.93	212.17
$N = 3000, P = 5, T = 5$	29.66	1.90	123.47	2.34	21.13
$N = 3000, P = 5, T = 20$	36.73	3.43	169.27	11.92	341.99

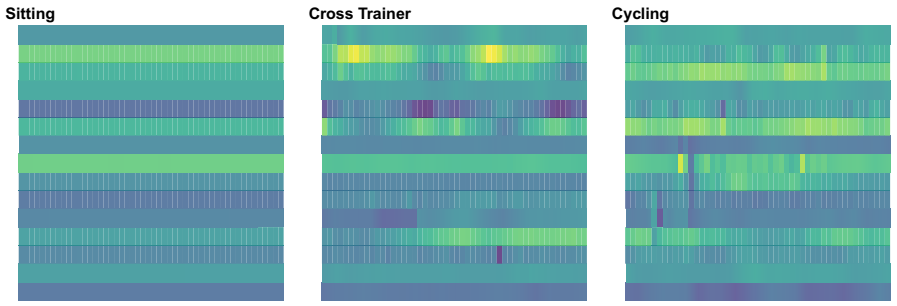

Fig. 6 Graphical representation of one observation from each of the three cluster in the activity tracking dataset. Cf. Fig. 2

and the sides of the knees of 8 subjects at the frequency of 25 measurements per second for 300 seconds. For each sensor location, nine variables have been recorded: the x,y,z axes acceleration, the x,y,z axes rate of turn, and the x,y,z axes Earth's magnetic field. Data are publicly available[1] and have been extensively described by Altun et al. (2010), Altun and Barshan (2010), Barshan and Yüksek (2014).

For the sake of illustration, we restrict the analysis on the 3 features detected by the accelerometer of the 5 sensors in 3 different activities performed by one subject only. The selected activities - sitting, exercising on a cross trainer and cycling on an exercise bike - are characterized by a variety of different degree of muscular activation and force produced. Each activity has been split in 150 sub-activities of 2 seconds, hence described by 50 measurements per variable. The resulting data set is then formed by 450 observations of dimension $P = 15$ and $T = 50$, grouped in three classes of activities.

Figure 6 illustrates an example of individual observation for each of the three activity. The goal of the analysis is to identify the measurements pertaining to the same activity. After standardizing the data, we run modal clustering based on the a k-NN balloon estimator, with $k = 5\sqrt{N}$, consistently with the indications drawn from the simulations. For comparison, we also consider the partitions detected by model-based clustering built on a mixture of Normal matrix-variate distributions and by K-means. The number of clusters has been selected to maximize the BIC and, respectively, the Silhouette score.

Results, reported in Table 2, show a general accuracy of all the considered methods at disclosing differences among the activities. However, while modal clustering correctly identify three groups, with just a very small amount of misclassified observations, the two competitors tend to oversegment the data, so that the actual clusters are in fact partitioned into a number smaller subgroups.

5.2 COVID-19 outgrowth across countries

At the time of writing this paper, the whole world has been severely harmed by the COVID-19 virus, a pandemic globally causing the largest social and economic dis-

[1] https://archive.ics.uci.edu/ml/datasets/Daily+and+Sports+Activities.

Table 2 *Tracker activity* data: comparison among partitions detected by the three clustering methods

Sitting cross trainer cycling									
Modal clustering			K-means						
1	2	3	1	2	3	4	5	6	7
150	0	0	0	0	0	58	64	0	28
4	146	0	0	0	83	0	0	67	0
9	0	141	67	83	0	0	0	0	0
FM = 0.942			FM = 0.624						

Sitting cross trainer cycling						
Mixture of matrix variate Gaussians						
1	2	3	4	5	6	7
150	0	0	0	0	0	0
0	31	44	41	34	0	0
0	0	0	0	0	24	126
FM = 0.812						

ruption since the last century. To reduce the spreading of the virus, most of countries have implemented measures of quarantine and social distancing practices, canceled or postponed most of sport, religious, political, and cultural events, interrupted business and educational activities.

Being the virus still in action, and the overall situation still evolving, drawing general conclusions on its impact is currently not possible, also due to different information which the countries have gathered and relayed about it. It is anyway clear that the spreading and the evolution of the pandemic, as well as its impact in relation to the adopted control measures, have not been the same all over the world. With this respect, the goal of this application is to evaluate differences and similarities among the countries. However, to prevent inconsistencies among countries due to different states and reactions to the pandemic, we restrict the analysis to the *first wave*, i.e. the period February-June 2020.

The data we consider have been collected by the Oxford COVID-19 Government Response Tracker (OxCGRT, Hale et al. 2020) and refer to daily observations of the number of confirmed cases of COVID-19 in each country, the number of confirmed deaths, along with several indicators reflecting the level of government action on health policies, economic support, strictness of lockdown policies. Considered that many of these indicators are highly correlated but not always available for all the countries, our analysis accounts for just one of them, namely the Stringency index, which ranges from 0 to 100 and summarizes the government response measures to the pandemic in terms of schools and work spaces closing, cancellation of public events and gatherings, "shelter-in-place" orders, movement restrictions and the presence of informative and awareness campaigns.

The resulting data set has been integrated with some further variables, intended to provide a rough indication about the economy and the demography of the countries. Specifically, the annual GDP based on purchasing power parity of each country is

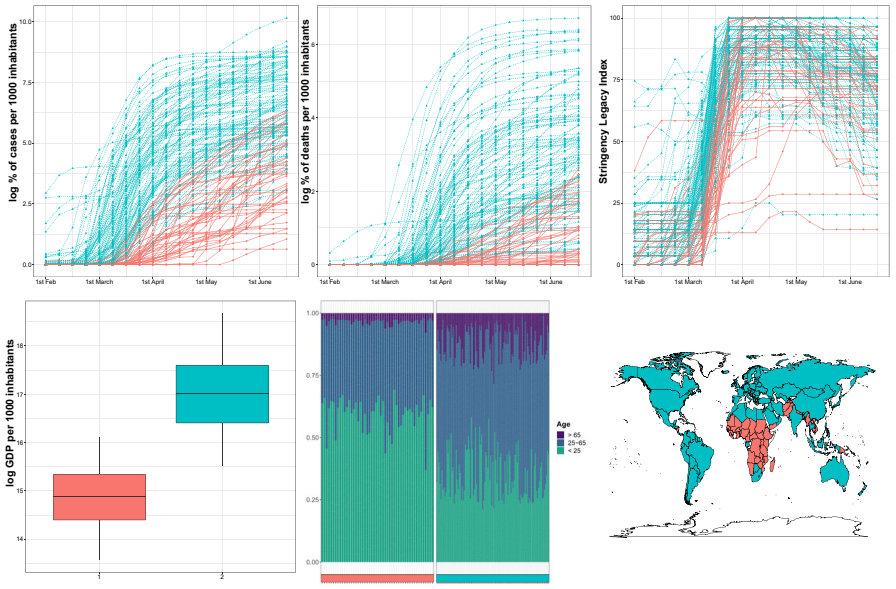

Fig. 7 Identified clusters for the considered variables in the COVID-19 study. The top panels shows the groups difference in confirmed cases, confirmed deaths, and stringency index. The bottom left panel shows the difference in GDP. The bottom middle panel shows the age distribution in the two groups. The bottom right panel shows the geographical distribution of the two identified groups

considered[2], as well as the population size and the age distribution[3], grouped in the three classes of population younger than 25, from 25 to 65, and over 65 years old.

Data have been pre-processed as follows: the logarithm of the number of confirmed cases and deaths per 1000 inhabitants and of the GDP per 1000 inhabitants have been evaluated, along with the percentage of population for each of the three age classes. The first two variables and the Stringency index, observed on a daily basis, have been averaged to get a weekly frequency, ranging from February, 1 to June, 15. The final individual observation is a matrix with dimension $P = 7$ and $T = 19$. Since GDP and age distribution refer to a yearly basis, their value has been kept constant over the 19 considered weeks. All the variables have been afterwards standardized. A few countries have been removed from the analysis, due to the presence of missing values, thus resulting in a final sample of size $N = 161$ countries.

Similarly to the Activity Tracking example, modal clustering has been run based on a k-nearest neighbors balloon estimator, with $k = 5\sqrt{N}$. Results, illustrated in Fig. 7, show an interesting pattern emerging from the data, with two clusters of countries having rather distinct characteristics. The largest group gathers all the countries over Europe, almost the entire America and Oceania, and many Asiatic countries, while the other group covers most of Africa and a few countries from Asia. The latter cluster is the one which the pandemic has harmed less severely in terms of both cases and deaths. While the answer of the governments, as measured by the stringency index,

[2] https://data.worldbank.org/indicator/NY.GDP.MKTP.CD.

[3] https://population.un.org/wpp/Download/Standard/Interpolated.

has not been in general weaker than in the countries assigned to the other cluster, the intervention in these countries has been in most of cases delayed, coherently with a lower perceived risk due to the limited virus spreading. Compared with the largest group, these countries have a demographic structure characterized by a larger proportion of young people, and a lower proportion of older people. This is consistent with the known behaviour of the COVID-19 virus, distressing especially older people. This apparently counterintuitive result, which labels the undeveloped countries as the less impacted by the pandemic, has in fact rather sound motivations. On one hand, the general economic and health conditions of the undeveloped countries have likely prevented accurate testing and tracking policies, so that we shall live with a limited reliability of the data. On the other hand, the social and demographic characteristics of the undeveloped countries have likely contribute to weaken the spreading of the pandemic, due to the generally young age, a prevalent family care of older people, and a limited mobility to and from outside the country.

It is worth noting that we ran a similar analysis also on the subset of European countries, for which further information is available (e.g. number of hospital beds per 1000 inhabitants and life expectation). European countries, taken on their own, split in two clusters, essentially formed by eastern and central Europe, and whose interpretation does not substantially depart from the one given for the whole word.

6 Discussion

Due to its unsupervised nature, clustering is a difficult task. The lack of an undisguised ground truth to pursue motivates a large use of visual inspection tools to get a sense of possible patterns in the data. However, high dimensionality may prevent graphical exploration to be actually fruitful, since only incomplete descriptions of the data are possible. Most of clustering methods are severely challenged in this framework. Distance-based methods, for instance, rely on the use of heuristic criteria for determining the number of clusters; on the other hand, model-based clustering requires awkward to verify assumptions on the cluster shape. The scarce reliability of visual inspection tools then turns out to be rather limiting when using such approaches.

Due to a reference cluster concept not constrained to any specific shape and to a determination of the number of groups as an integral part of the estimation procedure, modal clustering can be in principle applied even when an informative visual exploration of the data is prevented, as it may occur with matrix-variate data. In this work we have discussed how this approach can be extended to three way data structures, and faced the problem both with respect to the issue of density estimation and the one of mode detection. It is worth to note that most of the theory developed for the multivariate modal clustering framework is in principle generalizable to the matrix-variate case. The space $M_{P\times T}$ of $P \times T$ matrices is indeed a vector space, which hence maintains the standard properties of sum and multiplication by scalar and can be naturally equipped with a norm. This naturally allows for extending the results from Morse theory (see, for instance, Chacón 2015), other properties of kernels density estimators and their derivatives than those hereby discussed (e.g. Chacón and Duong 2018), as well as convergence results for the mean-shift (Arias-Castro et al. 2016). Leveraging

on these generalizations, it is further possible to derive extensions of inferential procedures to the matrix-variate framework. For example, the recent works by Duong et al. (2008), Genovese et al. (2016) and Ferraccioli and Menardi (2021) discuss the tricky task of testing mode significance, whose extension to the matrix-variate setting would be especially useful.

In fact, even when such generalizations are theoretically possible, they might not have a considerable significance. Most of the theory developed for multivariate modal clustering relies on asymptotics. In high dimensional spaces, where matrix-valued data turn out to be intrinsically defined, the sample size should ideally reach unfeasible magnitude in order to make such results hold. Also, required assumptions that appear fairly reasonable in small to moderate dimensions, become quite difficult to check in the space of matrices. For example, it is not uncommon to assume that the observations lie in a subspace, even nonlinear. This can possibly lead to a degenerate Hessian, which would further increase the complexity in the theoretical analysis.

On the other hand, numerical explorations performed in this work have shown that the situation, in practice, is not that critical as it might in principle appear. Since cluster separation tends to increase with the data dimensionality, the gross clustering structure is usually identified. While often the problem of spurious clusters cannot be straightened out completely, modal clustering has proven extraordinarily effective even when the matrix overall dimension is in the order of several hundreds and exceeds the sample size. Adaptive tools which account for the local characteristics of the data have proven to be quite effective in this context. Defining the amount of (local) smoothing, here intended as the proportion of sample neighbors to account for, is still an open problem. In our exploration we have considered simple heuristic criteria, highlighting that a large amount of smoothing is usually advisable, especially when the matrix dimension is large. However, defining more rigorous criteria targeted to the specific problem would be desirable. Additionally, while resorting to such local solutions has shown to be convenient to contain the curse of dimensionality, we shall live with the challenges arising from a non trivial extension of the inherent theory, which is left for future work.

Funding Open access funding provided by Università degli Studi di Padova within the CRUI-CARE Agreement.

Appendix A: Proof of Proposition 1

To establish the expression for the AMISE in Proposition 1, we start from the standard decomposition (3) and consider its characterization in terms of asymptotic IV and ISB.

Let us start by analyzing the asymptotic behaviour of the bias term. With a change of variables, the expected value of $\hat{f}(X;h)$ may be expressed as

$$\begin{aligned}\mathbb{E}(\hat{f}(X;h)) &= \int_{\mathbb{R}^{P,T}} h^{-P\cdot T} K(h^{-1}(X-Y)) f(Y) dY \\ &= \int_{\mathbb{R}^{P,T}} K(Z) f(X-hZ) dZ. \end{aligned} \tag{9}$$

The term $f(X-hZ)$ may be approximated with a Taylor expansion around X, thus obtaining

$$f(X-hZ) = f(X) - h\,\mathrm{tr}(\nabla f(X)^\top Z) + \frac{1}{2}h^2 \mathrm{tr}(\mathsf{H} f(X)^\top Z \otimes Z) + o(h^2),$$

where H denotes the Hessian matrix of dimensions $PT \times PT$. Using the fact that $\int_{\mathbb{R}^{P,T}} K(X)dX = 1$ and that $\int_{\mathbb{R}^{P,T}} XK(X)dX = 0$, and plugging the Taylor expansion into Eq. (9), we get

$$\begin{aligned}\mathbb{E}(\hat{f}(X;h)) &= f(X) + \frac{1}{2}h^2 \mathrm{tr}(\mathsf{H} f(X)^\top m_2(K)\mathbb{I}_{P\cdot T}) + o(h^2) \\ &= f(X) + \frac{1}{2}h^2 m_2(K)\mathrm{tr}(\mathsf{H} f(X)) + o(h^2) \\ &= f(X) + \frac{1}{2}h^2 m_2(K)\Delta f + o(h^2).\end{aligned}$$

Hence the approximated squared bias is

$$[\mathbb{E}(\hat{f}(X;h)) - f(X)]^2 = \frac{1}{4}h^4 m_2^2(K)\Delta f^2 + o(h^4).$$

By integrating with respect to X we obtain

$$\mathrm{ISB}(\hat{f}(\cdot;h)) = \frac{1}{4}h^4 m_2(K)^2 R(\Delta f) + o(h^4). \tag{10}$$

The variance of $\hat{f}$ is given by

$$\begin{aligned}\mathrm{Var}(\hat{f}(X;h)) = N^{-1} \int_{\mathbb{R}^{P,T}} h^{-2P\cdot T} K(h^{-1}(X-Y))^2 f(Y) dY - \\ N^{-1} \left(\int_{\mathbb{R}^{P,T}} h^{-P\cdot T} K(h^{-1}(X-Y)) f(Y) dY \right)^2. \end{aligned} \tag{11}$$

Starting from the first term in (11), and integrating it with respect to X, we obtain

$$N^{-1} \int_{\mathbb{R}^{P,T}} \int_{\mathbb{R}^{P,T}} h^{-2P\cdot T} K(h^{-1}(X-Y))^2 f(Y) dY dX$$

$$= N^{-1}h^{-(P\cdot T)} \int_{\mathbb{R}^{P,T}} \int_{\mathbb{R}^{P,T}} K(Z)^2 f(X - hZ) dZ dX$$
$$= N^{-1}h^{-(P\cdot T)} R(K), \tag{12}$$

where the first equality follows from the change of variable $Z = h(X - Y)$ and the second one from Fubini's theorem. For the second term in Eq. (11) we can take advantage of the previous calculations for $\mathbb{E}(\hat{f}(X; h))$ to obtain

$$N^{-1} \int_{\mathbb{R}^{P,T}} \int_{\mathbb{R}^{P,T}} h^{-2P\cdot T} K(h^{-1}(X - Y))^2 f(Y) dY dX$$
$$= N^{-1}R(f) + o(N^{-1}).$$

Given the assumption (iii), and in view of Eq. (12), it follows that this second term in the IV is of a smaller order than the first one. Therefore,

$$\mathrm{IV}(\hat{f}(\cdot; h)) = N^{-1}h^{-(P\cdot T)} R(K) + o(N^{-1}h^{-(P\cdot T)}). \tag{13}$$

Combining Eqs. (10) and (13), it follows that an asymptotic approximation to the MISE can be written as

$$\mathrm{AMISE}(\hat{f}(\cdot; h)) = N^{-1}h^{-(P\cdot T)} R(K) + \frac{1}{4}h^4 m_2(K)^2 R(\Delta f).$$

The optimal bandwidth (4) is derived via minimization of the AMISE, by identifying the root of

$$\frac{\partial}{\partial h}\mathrm{AMISE}(\hat{f}(\cdot; h)) = -\frac{(P \cdot T)R(K)}{Nh^{-(P\cdot T)-1}} + h^3 m_2(K)^2 R(\Delta f) = 0.$$

Appendix B: Proof of Proposition 2

Since the maxima of a function f satisfies $\nabla f = 0$, a standard formulation of a gradient ascent algorithm on its estimate s the following:

$$Y^{(s+1)} = Y^{(s)} + \alpha \nabla \hat{f}(Y^{(s)}). \tag{14}$$

For the specific case (2), simple differentiation rules of matrix-variate function lead to

$$\nabla \hat{f}(X; h) = \frac{1}{Nh^{P\cdot T}} \sum_{n=1}^{N} \nabla K \left(h^{-1}(X_n - X) \right). \tag{15}$$

The use of a spherically symmetric kernel K allows recasting to the simpler use of a function of real variable $\kappa : \mathbb{R}_+ \mapsto \mathbb{R}$, known as *profile* of K, via the representation

$$K(X) = \frac{1}{2}\kappa(||X||_F^2) = \frac{1}{2}\kappa(\mathrm{tr}(X^\top X)),$$

where $||\cdot||_F$ is the Frobenius norm. Note that the first equality highlights the similar structure of the matrix-variate kernel to the standard multivariate kernel, where the Euclidean norm replaces the Frobenius norm. Hence, the (15) turns into the following:

$$\begin{aligned}\nabla \hat{f}(X;h) &= \frac{1}{N}\frac{1}{2h^{P\cdot T}}\sum_{n=1}^{N}\nabla\kappa\left(h^{-2}\mathrm{tr}((X_n-X)^\top(X_n-X))\right)\\ &= -\frac{1}{N}\frac{1}{h^{P\cdot T+2}}\sum_{n=1}^{N}\kappa'\left(h^{-2}\mathrm{tr}((X_n-X)^\top(X_n-X))\right)(X_n-X). \quad (16)\end{aligned}$$

Replacing the (16) in the $\nabla\hat{f}$ term of (14), we obtain:

$$\begin{aligned}Y^{(s+1)} &= Y^{(s)} - \alpha\frac{1}{Nh^{P\cdot T+2}}\sum_{n=1}^{N}\kappa'\left(h^{-2}\mathrm{tr}((X_n-X)^\top(X_n-Y^{(s)}))\right)(X_n-Y^{(s)})\\ &= Y^{(s)} - \alpha\left[\frac{1}{N}\frac{1}{h^{P\cdot T+2}}\sum_{n=1}^{N}\kappa'\left(h^{-2}\mathrm{tr}((X_n-X)^\top(X_n-Y^{(s)}))\right)X_n+\right.\\ &\quad \left.\frac{1}{N}\frac{1}{h^{P\cdot T+2}}\sum_{n=1}^{N}\kappa'\left(h^{-2}\mathrm{tr}((X_n-X)^\top(X_n-Y^{(s)}))\right)Y^{(s)}\right], \quad (17)\end{aligned}$$

and setting an adaptive step size

$$\alpha = \alpha_s = \left[-\frac{1}{N}\frac{1}{h^{P\cdot T+2}}\sum_{n=1}^{N}\kappa'\left(h^{-2}\mathrm{tr}((X_n-Y^{(s)})^\top(X_n-Y^{(s)}))\right)\right]^{-1}.$$

we obtain the thesis.

References

Altun K, Barshan B (2010) Human activity recognition using inertial/magnetic sensor units. In: International workshop on human behavior understanding. Springer, Berlin, pp 38–51

Altun K, Barshan B, Tunçel O (2010) Comparative study on classifying human activities with miniature inertial and magnetic sensors. Pattern Recogn 43(10):3605–3620

Arias-Castro E, Mason D, Pelletier B (2016) On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. J Mach Learn Res 17(1):1487–1514

Barshan B, Yüksek MC (2014) Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. Comput J 57(11):1649–1667

Basford KE, McLachlan GJ (1985) The mixture method of clustering applied to three-way data. J Classif 2(1):109–125

Caro-Lopera FJ, Farías GG, Balakrishnan N (2016) Matrix-variate distribution theory under elliptical models-4: joint distribution of latent roots of covariance matrix and the largest and smallest latent roots. J Multivar Anal 145:224–235

Chacón JE (2015) A population background for nonparametric density-based clustering. Stat Sci 30(4):518–532

Chacón JE, Duong T (2018) Multivariate kernel smoothing and its applications. CRC Press, Cambridge

Chakraborty R, Vemuri BC et al (2019) Statistics on the Stiefel manifold: theory and applications. Ann Stat 47(1):415–438

Diggle P, Diggle PJ, Heagerty P, Liang K-Y, Heagerty PJ, Zeger S et al (2002) Analysis of longitudinal data. Oxford University Press, Oxford

Ding S, Cook DR (2018) Matrix variate regressions and envelope models. J R Stat Soc Ser B (Stat Methodol) 80(2):387–408
Dryden IL, Koloydenko A, Zhou D et al (2009) Non-euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann Appl Stat 3(3):1102–1123
Duong T (2019) ks: Kernel Smoothing. R package version 1.11.5. https://CRAN.R-project.org/package=ks
Duong T, Cowling A, Koch I, Wand MP (2008) Feature significance for multivariate kernel density estimation. Comput Stat Data Anal 52(9):4225–4242
Duong T, Beck G, Azzag H, Lebbah M (2016) Nearest neighbour estimators of density derivatives, with application to mean shift clustering. Pattern Recogn Lett 80:224–230
Ferraccioli F, Menardi G (2021) A nonparametric test for mode significance. In: Porzio G, Rampichini C, Bocci C (eds) Cladag, 2021, Book of abstracts and short papers. Firenze University Press, New York, pp 388–391
Fukunaga K, Hostetler L (1975) The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory 21(1):32–40
Gallaugher MP, McNicholas PD (2018) Finite mixtures of skewed matrix variate distributions. Pattern Recogn 80:83–93
Genovese CR, Perone-Pacifico M, Verdinelli I, Wasserman L (2016) Non-parametric inference for density modes. J R Stat Soc B 78:99–126
Ghassabeh YA (2015) A sufficient condition for the convergence of the mean shift algorithm with gaussian kernel. J Multivar Anal 135:1–10
Gupta AK, Nagar DK (2018) Matrix variate distributions, vol 104. CRC Press, Cambridge
Hale T, Webster S, Petherick A, Phillips T, Kira B (2020) Oxford covid-19 government response tracker
Hennig C, Meila M, Murtagh F, Rocci R (2015) Handbook of cluster analysis. CRC Press, Cambridge
Jacques J, Preda C (2014) Model-based clustering for multivariate functional data. Comput Stat Data Anal 71:92–106
Kroonenberg PM (2008) Applied multiway data analysis, vol 702. Wiley, New York
Makhoul J (1980) A fast cosine transform in one and two dimensions. IEEE Trans Acoust Speech Signal Process 28(1):27–34
Menardi G (2016) A review on modal clustering. Int Stat Rev 84(3):413–433
R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. https://www.R-project.org/
Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math 20:53–65
Sakata T (2016) Applied matrix and tensor variate data analysis. Springer, Berlin
Sarkar S, Zhu X, Melnykov V, Ingrassia S (2020) On parsimonious models for modeling matrix data. Comput Stat Data Anal 142:106822
Schmutz A, Jacques J, Bouveyron C, Cheze L, Martin P (2020) Clustering multivariate functional data in group-specific functional subspaces. Comput Stat 1–31
Strang G (1999) The discrete cosine transform. SIAM Rev 41(1):135–147
Stuetzle W (2003) Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. J Classif 20(1):25–47
Tomarchio SD, Punzo A, Bagnato L (2020) Two new matrix-variate distributions with application in model-based clustering. Comput Stat Data Anal 152:107050
Vermunt JK (2007) A hierarchical mixture model for clustering three-way data sets. Comput Stat Data Anal 51(11):5368–5376
Vichi M, Rocci R, Kiers HA (2007) Simultaneous component and clustering models for three-way data: within and between approaches. J Classif 24(1):71–98
Viroli C (2011) Finite mixtures of matrix normal distributions for classifying three-way data. Stat Comput 21(4):511–522
Viroli C (2012) On matrix-variate regression analysis. J Multivar Anal 111:296–309
Viroli C et al (2011) Model based clustering for three-way data structures. Bayesian Anal 6(4):573–602
Wang M, Fischer J, Song YS et al (2019) Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition. Annals Appl Stat 13(2):1103–1127
Zhou H, Li L (2014) Regularized matrix regression. J R Stat Soc Ser B (Stat Methodol) 76(2):463–483