The Guedon-Vershynin Semi-Definite Programming approach to low dimensional embedding for unsupervised clustering

Stephane Chretien, Clément Dombry, Adrien Faivre

A Semi-Definite Programming approach to low dimensional embedding for unsupervised clustering

Article *in* Frontiers in Applied Mathematics and Statistics · June 2016

DOI: 10.3389/fams.2019.00041

3 authors:

Stéphane Chrétien
Université Lumiere Lyon 2
103 PUBLICATIONS **454** CITATIONS
SEE PROFILE

Clement Dombry
University of Franche-Comté
62 PUBLICATIONS **481** CITATIONS
SEE PROFILE

Adrien Faivre
University of Franche-Comté
4 PUBLICATIONS **9** CITATIONS
SEE PROFILE

The Guedon-Vershynin Semi-Definite Programming approach to low dimensional embedding for unsupervised clustering

Stéphane Chrétien * Clément Dombry † Adrien Faivre ‡

July 3, 2019

Abstract

This paper proposes a method for estimating the cluster matrix in the Gaussian mixture framework via Semi-Definite Programming. Theoretical error bounds are provided and a (non linear) low dimensional embedding of the data is deduced from the cluster matrix estimate. The method and its analysis is inspired by the work by Guédon and Vershynin on community detection in the stochastic block model. The adaptation is non trivial since the model is different and new Gaussian concentration arguments are needed. Our second contribution is a new Bregman-ADMM type algorithm for solving the semi-definite program and computing the embedding. This results in an efficient and scalable algorithm taking only the pairwise distances as input. The performance of the method is illustrated via Monte Carlo experiments and comparisons with other embeddings from the literature.

1 Introduction

Low dimensional embedding is a key to many modern data analytics. Data are better understood after choosing the best coordinates, i.e. embedding, and extracting the main features. Based on a compressed description, the data can then be projected, visualized or clustered more reliably and efficiently. The goal of the present paper is to present an efficient technique for joint embedding and clustering, based on pairwise affinity analysis and reliable convex optimisation.

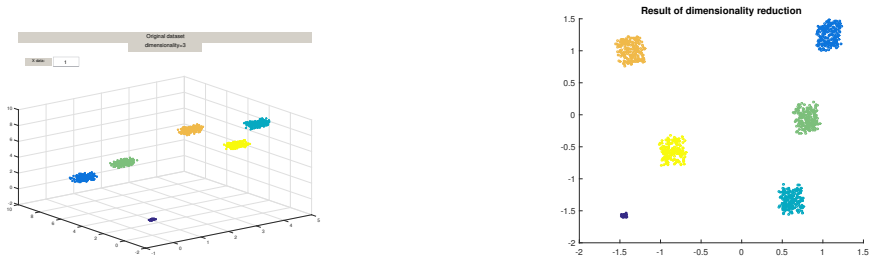
Combining the goals of reducing dimensionality and clustering in a principled manner is challenging and novel, but also draws on ideas from spectral clustering [25], [3], and Semi-Definite embedding, as in [15]. The main idea

*National Physical Laboratory, Hampton Road, TW11 0LW, Teddington, UK

†Laboratoire de Mathematiques de Besançon, Université de Franche-Comté, 16 route de Gray, 25030, Besançon, France

‡Digitalsurf, Besançon, France

behind such methods, is to approximately preserve the pairwise distances in the dataset, with the goal of discovering, via an appropriate coordinate change, the correct parametrisation of the potentially low dimensional non-linear manifold that essentially contains the data. An example of non-linear low dimensional embedding, such as Diffusion Maps, is shown in Figure 1.



(a) Original 3D Cluster

(b) Mapped data using Diffusion Maps

Figure 1: The mapping of a 3D cluster using Diffusion Maps from the Matlab package drtoolbox https://lvdmaaten.github.io/drtoolbox/

Apart from this previous works, standard clustering techniques usually start from already embedded data as obtained after e.g. PCA processing. Said otherwise embedding and clustering are often considered as completely separate tasks. Based on embedded data, mainstream clustering techniques are non-parametric (as K-means , K-means++, etc [14], [10]) and model based clustering (as Mixture Models [16]). These approaches often rely on non-convex optimisation such as Lloyd's algorithm and Expectation-Maximisation [8], [7]. However, even careful implementation of these algorithms leads to some uncertainty as to whether one has finally obtained a relevant optimiser. In Mixture Model-based clustering, one also faces the issue of degeneracy [5]. As a result, a growing body of researched has emerged lately concerning the study of convex relaxation for the clustering problem [2], [18], etc. Other approaches such as ClusterPath [13] have also been proposed; see [23], [22] and [27] for interesting extensions. One drawback of theses approaches is the presence of many hyperparameters without robust rules to select them. The present work pursues this recent trend of promoting convex optimisation-based clustering based on low rank cluster matrix estimation, as a way of ensuring that no spurious local optimiser is used in the process of clustering-based decision making in data analytic studies.

Our starting point in this attempt at finding appropriate embeddings for clustering is the method by Guedon and Vershynin [11], initially developed for community detection in the stochastic block model (SBM). In mathematical terms, the SBM considers a random graph based on a set of vertices partitioned into clusters and with random edges between vertices, all edges are independent

and the probabilities of edges depend only on the cluster structure. The usual assumption made in the SBM is that the probabilities are larger within clusters than across clusters. The problem of recovering clusters from a random empirical adjacency matrix has been a topic of extensive research, triggered by the work of McSherry [17] and which quickly developed into a beautiful body of impressive results and achievements; see Abbe et al. [1], Heimlicher et al. [12], Mossel et al. [19], [20], [21]. Guedon and Vershynin recently showed that the cluster matrix can be estimated via Semi-Definite Programming (SDP) with an explicit control of the error rate.

From a technical perspective, our contribution is three fold.

- First, we generalise the Guedon/Vershynin approach in order to deal with the Gaussian Cluster Model (GCM) and show that the cluster matrix in the GCM can also be estimated by solving an SDP. For doing so, we use an affinity matrix as input that depends only on the pairwise distances between observations. Contrarily to the adjacency matrix arising in the SBM, our affinity matrix from the GCM has non independent entries, thus making the analysis non trivial.

- Our second contribution is to demonstrate in practice that the estimated cluster matrix yields a natural associated embedding. Indeed, quite similarly to spectral clustering, the eigenvectors of the estimated cluster matrix provide a meaningful embedding. Contrarily to standard embedding methods such as PCA, Laplacian eigenmaps, Maximum Variance Unfolding, t-SNE, etc, the embedding does not try to preserve pairwise distances but rather to estimate the cluster matrix. The intuition for using the cluster matrix is supported by Remark 1.6 in [11] which we now quote: *It may be convenient to view the cluster matrix as the adjacency matrix of the cluster graph, in which all vertices within each community are connected and there are no connections across the communities. This way, the Semi-Definite program takes a sparse graph as an input, and it returns an estimate of the cluster graph as an output. The effect of the program is thus to "densify" the network inside the communities and "sparsify" it across the communities.*

- Our third contribution is to propose a new scalable algorithm for solving the main Semi-Definite Programming problem at the heart of [11] and our approach to embedding and clustering. Our new method is based on a linearised version of the Alternating Direction Method of Multipliers (ADMM) together with a pragmatic implementation of the constraints, that allows us the avoid solving the original Semi-Definite Program via interior point methods.

The paper is organized as follows. The SDP approach for estimating the cluster matrix, the associated embedding and the main theoretical results are presented in Section 2. The proofs are postponed to Section A in the appendix.

The algorithmic considerations for the resolution of the SDP are discussed in the supplementary material, where an efficient algorithm is described as well as a practical method for selecting the unknown tuning parameter. Section 3 is devoted to the presentation of simulation results, demonstrating the potential of the proposed method. Some technical background on Gaussian concentration and Grothendieck inequality is provided in Appendices.

2 Main results

2.1 Framework: the Gaussian Cluster Model

The mathematical framework is the following. We assume that we observe a data set $x_1, \dots, x_n \in \mathbb{R}^d$ over a population of size n. The population is partitioned into K clusters $\mathcal{C}_1, \dots, \mathcal{C}_K$ of size $n_1, \dots, n_K$ respectively, i.e. $n = n_1 + \cdots + n_K$. We assume the standard Gaussian Cluster Model for the data: the observations x_i are independent with

$$x_i \sim \mathcal{N}(\mu_k, \Sigma_k) \quad \text{if } i \in \mathcal{C}_k \tag{1}$$

with $\mu_k \in \mathbb{R}^d$ the cluster mean and $\Sigma_k \in \mathbb{R}^{d\times d}$ the cluster covariance matrix. The Gaussian Mixture model specifies the addtional information about the probabilities of belonging to each cluster and the Gaussian Cluster Model corresponds to conditioning the Gaussian Mixture Model on the values of the latent cluster indicator variables [16] [1].

The clustering problem aims at recovering the clusters $\mathcal{C}_k$, $1 \leqslant k \leqslant K$, based on the data x_i, $1 \leqslant i \leqslant n$, only. For each $i = 1, \dots, n$, we will denote by k_i the index of the cluster to which i belongs. The notation $i \sim j$ will mean that i and j belong to the same cluster. The cluster matrix $\bar{Z}$ is the $n \times n$ matrix defined by

$$\bar{Z}_{i,j} = \begin{cases} 1 & \text{if } i \sim j \\ 0 & \text{otherwise} \end{cases}, \quad 1 \leqslant i, j \leqslant n. \tag{2}$$

It determines entirely the clusters and, up to a reordering of the points, it is a block-diagonal matrix with a block of ones for each cluster.

Note that the Gaussian Cluster Model slightly differs from the usual Gaussian mixture model where the data set consists in independent observations from the Gaussian mixture $\sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mu_k, \Sigma_k)$, with $(\pi_k)_{1\leqslant k\leqslant K}$ the mixture distribution. In the Gaussian mixture model, the cluster sizes $(n_1, \dots, n_k)$ are random with multinomial distribution of size n and probability parameters $(\pi_1, \dots, \pi_K)$.

[1] Extending our study to the setting of Gaussian Mixture Models using the relationship with the Gaussian Cluster Model based on this conditioning is a somewhat tedious but not difficult task.

2.2 Dimensionality reduction

As proved in [4] the data can be projected onto a space of dimension $\log(K)$ while still preserving separation when the data are assumed to belong to separated ellipsoids. Therefore, we can consider in the rest of the paper that d is of the order of $\log(K)$. The results of our main theorem below will give the appropriate scaling of the parameters that will ensure the well separatedness of the data.

2.3 Embedding associated with the estimated cluster matrix

We will define in the next section an estimate $\widehat{Z}$ of the cluster matrix $\bar{Z}$. We discuss now how a low dimensional embedding of the data set can be associated with the estimate $\widehat{Z}$. The main idea is to use the fact that the cluster matrix $\bar{Z}$ defined by (2) has a very specific eigenstructure: denoting by $\mathcal{C}_1, \ldots, \mathcal{C}_K$ the index set of each cluster and by $1_{\mathcal{C}_k} \in \{0,1\}^n$ the indicator vector of cluster $\mathcal{C}_k$, we have

$$\bar{Z} = \sum_{k=1}^{K} 1_{\mathcal{C}_k} 1_{\mathcal{C}_k}^t$$

and we deduce that

- the rank of $\bar{Z}$ is K,
- the nonzero eigenvalues of $\bar{Z}$ are $|\mathcal{C}_1|, \ldots, |\mathcal{C}_K|$ with associated eigenvectors $1_{\mathcal{C}_1}/\sqrt{|\mathcal{C}_1|}, \ldots, 1_{\mathcal{C}_K}/\sqrt{|\mathcal{C}_1|}$.

We assume in the sequel that the cluster sizes are all different so that all non-zero eigenvalues have multiplicity one. The clusters can hence be recovered from the eigenstructure of the matrix $\bar{Z}$: the label of the sample point x_i is the index of the only eigenvector whose i-th component is non zero. Indeed, all other eigenvectors associated with a non-zero eigenvalue have i-th component equal to zero.

The estimate $\widehat{Z}$ of the matrix $\bar{Z}$ can be used in practice to embed the data into the space $\mathbb{R}^K$ by associating each data point x_i to the vector consisting of the i-th coordinate of the K first eigenvectors of $\widehat{Z}$. Given this embedding, if we can prove that $\widehat{Z}$ accurately estimates the cluster matrix $\bar{Z}$, one can then apply any clustering method of choice to recover the clustering pattern of the original data. The next section gives a method for computing an estimator $\widehat{Z}$ of $\bar{Z}$ using the SDP approach by Guedon and Vershynin.

2.4 Guedon and Vershynin's Semi-Definite Program

We now turn to the estimation $\widehat{Z}$ of the cluster matrix using Guedon and Vershynin's Semi-Definite Programming based approach. Whereas Vershynin and Guédon [11] were interested in analyzing the Stochastic Block Model for community detection, we propose a study of the Gaussian Cluster Model and therefore

prove that their approach has a great potential applicability in embedding of general data sets beyond the Stochastic Block Model setting.

Based on the data set $x_1, \dots, x_n$, we construct the affinity matrix A by

$$A = \big(f(\|x_i - x_j\|_2)\big)_{1 \leqslant i,j \leqslant n} \tag{3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm on $\mathbb{R}^d$ and $f : [0, +\infty) \to [0, 1]$ a non-increasing affinity function. A popular choice is the Gaussian affinity

$$f(h) = e^{-(h/h_0)^2}, \quad h \geqslant 0, \tag{4}$$

with $h_0 > 0$, and other possibilities are

$$f(h) = e^{-(h/h_0)^a}, \quad f(h) = (1 + (h/h_0))^{-a},$$
$$f(h) = (1 + e^{h/h_0})^{-a} \quad \dots$$

Before stating the Semi-Definite Program, we introduce some matrix notations. The usual scalar product between matrices $A, B \in \mathbb{R}^{n \times n}$ is denoted by $\langle A, B \rangle = \sum_{1 \leqslant i,j \leqslant n} A_{ij} B_{ij}$. The notations $1_n \in \mathbb{R}^n$ and $1_{n \times n} \in \mathbb{R}^{n \times n}$ stand for the vector and matrices with all entries equal to 1. For a symmetric matrix $Z \in \mathbb{R}^{n \times n}$, the notation $Z \succeq 0$ means that Z the quadratic form associated to Z is non-negative while the notation $Z \geqslant 0$ means that all the entries of Z are non-negative.

With these notations, we define $\widehat{Z}$ as a solution of the Semi-Definite Program

$$\text{maximize } \langle A, Z \rangle \quad \text{subject to} \quad Z \in \mathcal{M}_{opt} \tag{5}$$

with $\mathcal{M}_{opt}$ the set of symmetric matrices $Z \in \mathbb{R}^{n \times n}$ such that

$$\begin{cases} Z \succeq 0 \\ Z \geqslant 0 \\ \operatorname{diag}(Z) = 1_n \\ \langle Z, 1_{n \times n} \rangle = \lambda_0 \end{cases} . \tag{6}$$

Here $\lambda_0 \in \mathbb{N}$ is the number of non-zero edges in the true cluster matrix and Guedon and Vershynin state in [11] that it can be estimated empirically. For further reference, note that a semi-definite positive matrix Z with non-negative entries and unit diagonal must have all entries in $[0, 1]$, so that $\mathcal{M}_{opt} \subset [0, 1]^{n \times n}$.

The heuristic justifying that $\widehat{Z}$ can be seen as an estimate of the cluster matrix $\bar{Z}$ is the following Lemma.

Lemma 2.1. *Consider the expected affinity matrix*

$$\bar{A} = \Big(\mathrm{E} f(\|x_i - x_j\|_2)\Big)_{1 \leqslant i,j \leqslant n}. \tag{7}$$

and assume

$$p = \inf_{i \sim j} \bar{A}_{i,j} > q = \sup_{i \not\sim j} \bar{A}_{i,j}. \tag{8}$$

Then, the cluster matrix $\bar{Z}$ defined by Eq. (2) *is the unique solution of*

$$\text{maximize} \quad \langle \bar{A}, Z \rangle \quad \text{subject to} \quad Z \in \mathcal{M}_{opt} \tag{9}$$

with $\lambda_0 = \sum_{k=1}^{K} n_k^2$.

The intuition behind condition (8) is that the average distance (or more precisely the average affinity) between two points within a same cluster is smaller than the average distance between two points from different clusters. This corresponds to the intuitive notion of clusters. Note that a similar condition appears in [11]. In the case of the Gaussian affinity function (4), we provide in Section 2.6 explicit formulas for the expected affinity matrix that can be used to check condition (8).

The SDP (5) appears as an approximation of the SDP (9) since the affinity matrix A can be seen as a noisy observation of the unobserved matrix $\bar{A}$. Concentration arguments together with Grothendieck theorem allow to prove that $A \approx \bar{A}$ in the sense of the ℓ^∞/ℓ^1-norm (see Proposition 2.4 below). In turn, this implies $\widehat{Z} \approx \bar{Z}$ in the sense of ℓ^1-norm in $\mathbb{R}^{n^2}$ (see Theorem 2.2 below) so that the SDP program (5) provides a good approximation $\widehat{Z}$ of the cluster matrix $\bar{Z}$. Note that in practice, λ_0 is unknown and must be estimated, see comment in section 5.1.5.

Proof of Lemma 2.1. This corresponds to Lemma 7.1 in [11]. □

2.5 Theoretical error bounds

Our main result is a non asymptotic upper bound for the probability that $\widehat{Z}$ differs from $\bar{Z}$ in ℓ^1-distance, that is an upper bound for

$$\left\| \widehat{Z} - \bar{Z} \right\|_1 = \sum_{1 \leqslant i,j \leqslant n} |\widehat{Z}_{i,j} - \bar{Z}_{i,j}|.$$

Theorem 2.2. *Consider the Gaussian Cluster Model* (1) *and assume that the affinity function f is ℓ-Lipschitz and that condition* (8) *is satisfied. Let*

$$t_0 = 8\sqrt{2 \log 2} K_G \sigma \ell / (p - q)$$

where $K_G \leqslant 1.8$ denotes the Grothendieck constant and $\sigma^2 = \frac{1}{n} \sum_{k=1}^{K} n_k \rho(\Sigma_k)$ with $\rho(\Sigma_k)$ the largest eigenvalue of the covariance matrix Σ_k.
Then, for all $t > t_0$,

$$\mathrm{P}\left(\left\| \widehat{Z} - \bar{Z} \right\|_1 > n^2 t \right) \leqslant 2 \exp\left(- \left(\frac{t - t_0}{c} \right)^2 n \right), \tag{10}$$

$$c = \frac{16\sqrt{2} K_G \ell \sigma}{p - q}. \tag{11}$$

Moreover, there exists a subset $\tau \subset \{1, \dots, n\}$ *with* $|\tau| \geqslant \frac{n}{2}$ *such that all* $t > t_0$,

$$\mathrm{P}\left(\left\| \left(\widehat{Z} - \bar{Z} \right)_{\tau \times \tau} \right\|_1 > nt \right) \leqslant 2 \exp\left(- \left(\frac{t - t_0}{c} \right)^2 n \right). \tag{12}$$

Proof. See Section A.1 in the Appendix. □

Theorem 2.2 has a simple consequence in terms of estimation error rate. After computing $\widehat{Z}$, it is natural to estimate the cluster graph $\bar{Z}$ by a random graph obtained by putting an edge between vertices i and j if $\widehat{Z}_{i,j} > 1/2$ and no edge otherwise. Then the proportion π_n of errors in the prediction of the $n(n-1)/2$ edges is given by

$$\begin{aligned} \pi_n &:= \frac{2}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} |1_{\{\widehat{Z}_{ij} > 1/2\}} - \bar{Z}_{ij}| \\ &\leqslant \frac{2}{n(n-1)} \sum_{1 \leqslant i < j \leqslant n} 2\, |\widehat{Z}_{ij} - \bar{Z}_{ij}| \\ &= \frac{2}{n(n-1)} \left\| \widehat{Z} - \bar{Z} \right\|_1 . \end{aligned}$$

The following corollary provides a simple bound for the asymptotic error.

Corollary 2.3. *We have almost surely*

$$\limsup_{n \to \infty} n^{-2} \left\| \widehat{Z} - \bar{Z} \right\|_1 \leqslant t_0 = \frac{8\sqrt{2 \log 2} K_G \sigma \ell}{p - q}.$$

In the case when the cluster means are pairwise different and fixed while the cluster variances converge to 0, i.e. $\sigma \to 0$, it is easily seen that the right hand side of the above inequality behaves as $O(\sigma)$ so that the error rate converges to 0. This reflects the fact that when all clusters concentrates around their means, clustering becomes trivial.

While our proof of Theorem 2.2 follows the ideas from Vershynin and Guédon [11], we need to introduce new tools to justify the approximation $A \approx \bar{A}$ in ℓ^∞/ℓ^1-norm. Indeed, unlike in the stochastic block model, the entries of the affinity matrix (3) are not independent. We use Gaussian concentration measure arguments to obtain the following concentration inequality. The ℓ^∞/ℓ^1 norm of a matrix $M \in \mathbb{R}^{n \times n}$ is defined by

$$\|M\|_{\infty \to 1} = \sup_{\|u\|_\infty \leqslant 1} \|Mu\|_1 = \max_{u,\ v \in \{-1,1\}^n} \sum_{i,j=1}^n u_i v_j M_{i,j}. \tag{13}$$

Proposition 2.4. *Consider the Gaussian cluster model* (1) *and assume the affinity function* f *is* ℓ*-Lipschitz. Then, for any* $t > 2\sqrt{2 \log 2}\ \ell\ \sigma$,

$$\begin{aligned} &\mathrm{P}\Big(\left\| A - \bar{A} \right\|_{\infty \to 1} > t\ n^2 \Big) \\ &\quad \leqslant 2 \exp\left(- \frac{\left(t - 2\sqrt{2 \log 2} \ell \sigma \right)^2}{32 \ell^2 \sigma^2}\ n \right). \end{aligned} \tag{14}$$

Proof. See Section A.2 in the Appendix. □

Theorem 2.2 assumes that λ_0 is known. It is worth noting that λ_0 corresponds to the number of edges in the cluster graph and that we can derive from the proof of Theorem 2.2 how the algorithm behaves when the cluster sizes are unknown, i.e. when the unknown parameter λ_0 is replaced with a different value λ. The intuition is given in Remark 1.6 in [11]: if $\lambda < \lambda_0$, the solution $\widehat{Z}$ will estimate a certain subgraph of the cluster graph with at most $\lambda_0 - \lambda$ missing edges; if $\lambda > \lambda_0$, the solution $\widehat{Z}$ will estimate a certain supergraph of the cluster graph with at most $\lambda - \lambda_0$ extra-edges.

2.6 Explicit formula for $\bar{A}$

In order to check condition (8), explicit formulas for the mean affinity matrix are useful. The next proposition solves the case of the Gaussian affinity function.

Proposition 2.5. *Assume that A is built using the Gaussian affinity function (4).*

- *Let i and j be in the same cluster $\mathcal{C}_k$. Then,*

$$\bar{A}_{i,j} = \prod_{l=1}^{d} \left(1 + 4(\sigma_{k,l}/h_0)^2\right)^{-1/2}$$

  with $(\sigma_{k,l}^2)_{1 \leqslant l \leqslant d}$ the eigenvalues of Σ_k.

- *Let i and j be in different clusters $\mathcal{C}_k$ and $\mathcal{C}_{k'}$. Then,*

$$\bar{A}_{i,j} = \prod_{l=1}^{d} \exp\left(-\frac{\langle \mu_k - \mu_{k'}, v_{k,k',l} \rangle^2}{h_0^2 + 2\sigma_{k,k',l}^2}\right) \left(1 + 2(\sigma_{k,k',l}/h_0)^2\right)^{-1/2}$$

  with $(\sigma_{k,k',l}^2)_{1 \leqslant l \leqslant d}$ and $(v_{k,k',l})_{1 \leqslant l \leqslant d}$ respectively the eigenvalues and eigenvectors of $\Sigma_k + \Sigma_{k'}$.

Proof. See Section A.3 in the Appendix. □

As an interesting consequence of Proposition 2.5, when the variance matrices from the Gaussian Cluster Model (1) are all equal and isotropic, that is $\Sigma_k = \sigma^2 \mathrm{Id}$ for all $k = 1, \ldots, K$ with $\sigma^2 > 0$, then the constants p and q from Equation (8) are given by

$$p = \left(1 + 4\sigma^2/h_0^2\right)^{-d/2}$$

and

$$q = \left(1 + 4\sigma^2/h_0^2\right)^{-d/2} \times \min_{1 \leqslant k \neq k' \leqslant K} \exp\left\{-\|\mu_{k'} - \mu_k\|^2/(h_0^2 + 4\sigma^2)\right\}.$$

Condition (8) is therefore satisfied (whatever the choice of $h_0 > 0$) as soon as the cluster means $(\mu_k)_{1 \leqslant k \leqslant K}$ are pairwise distinct which is a minimal identifiability condition. But of course the difference $p - q$ is an increasing function of the noise σ^2 and the bounds in Theorem 2.2 become looser for larger noise.

3 Simulation results

In all the experiments, the parameter h_0 in (4) was chosen as

$$h_0 = .5 * \max(\mathrm{diag}(X^t * X))^{1/2}.$$

The hyper-parameter λ was chosen so as to minimise the mean squared error between the estimated cluster matrix and the empirical affinity matrix.

3.1 Computing the actual clustering from the eigenvector coordinates

As for spectral clustering, the components of the most significant eigenvectors, i.e. the eigenvectors associated with the largest eigenvalues, are the coordinates of the embedded data. Given these embedded data, as advised in [26], the actual clustering can be computed using a minimum spanning tree method and removing the largest edges.

3.2 Comparison with standard embeddings on a 3D cluster example

Simulations have been conducted to assess the quality of the proposed embedding. In this subsection, we used the Matlab package drtoolbox [2] proposed by Laurens Van Maatten on a sample drawn from a 10 dimensional Gaussian Mixture Model with 4 components and equal proportions. In Figure 2, we show the original affinity matrix together with the estimated cluster matrix. In Figure 3, we compare the affinity matrix of data with the affinity matrix of the mapped data using various embeddings proposed in the drtoolbox package. This toy experiment shows that the embedding described in this paper can cluster as the same time as it embeds into a small dimensional subspace. This is not very surprising since our embedding is tailored for the joint clustering-dimensionality reduction purpose whereas most of the known existing embedding methods aren't. Given the fact that clustered data are ubiquitous in real world data analysis due to the omnipresence of stratified populations, taking the clustering purpose into account might be a considerable advantage.

3.3 Monte Carlo Experiments

In this section, we present some simulation experiments assessing the performance of the Guedon-Vershynin embedding for Gaussian Cluster Models.

[2] https://lvdmaaten.github.io/drtoolbox/

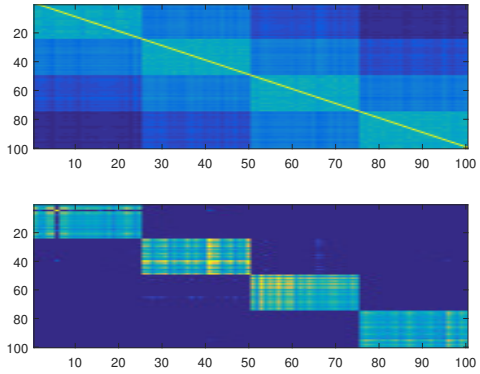


Figure 2: Original affinity matrix vs. Guedon Vershynin Cluster matrix

3.3.1 Setup

Our experiments were performed on problems of successive sample size 100, 200, ..., 1000 and number of clusters equal to 2, 5 and 10. The dimension of the Gaussian Mixture Model was set to 100. For each experiment, we performed 100 Monte Carlo repeats. All the results in this section show the average over the Monte Carlo experiments. Our Gaussian Cluster Model was built as follows: for a model with K clusters, we set the k^{th} component of each center to $10/3$, $20/3$, ..., $50/3$ for each cluster $k = 1, \ldots, K$. Then, the data are obtained by adding a unit variance i.i.d. Gaussian vector to the center of the cluster it belongs to. All clusters were taken to have equal size.

3.3.2 Selection of λ

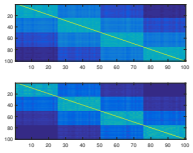
The value of λ was selected so as to minimise the Frobenius distance between $\hat{Z}$ and A. Model based selection rules will be discussed in a follow up paper.

3.3.3 Results

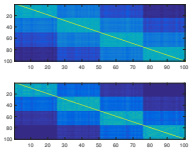
Figures 4 and 5 show the estimation error $\|\bar{Z} - \hat{Z}\|_1$ between the true and the estimated cluster matrix. These results illustrate Theorem 2.2 as they show that the error grows as a function of sample size. Moreover, the growth is quadratic as predicted by the theory of Section 2, and more precisely Eq. (10).

4 Conclusions

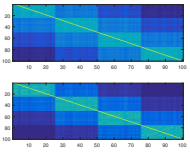
The goal of the present paper was to propose an analysis of Guedon and Vershynin's Semi-Definite Programming approach to the estimation of the cluster



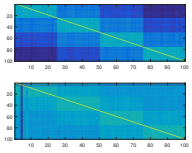
(a) Original affinity matrix vs. affinity matrix after PCA embedding



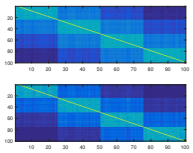
(b) Original affinity matrix vs. affinity matrix after MDS embedding



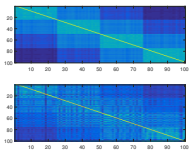
(c) Original affinity matrix vs. affinity matrix after Factor Analysis embedding



(d) Original affinity matrix vs. affinity matrix after t-SNE embedding



(e) Original affinity matrix vs. affinity matrix after Sammon embedding



(f) Original affinity matrix vs. affinity matrix after LLE embedding

Figure 3: The affinity matrix obtained after embedding using different methods from the Matlab package drtoolbox https://lvdmaaten.github.io/drtoolbox/

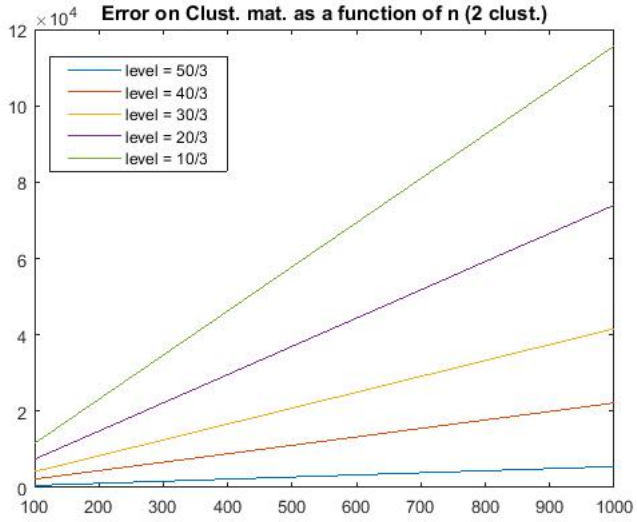


Figure 4: Estimation error $\|\bar{Z} - \hat{Z}\|_1$

matrix and show how this matrix can be used to produce an embedding for pre-

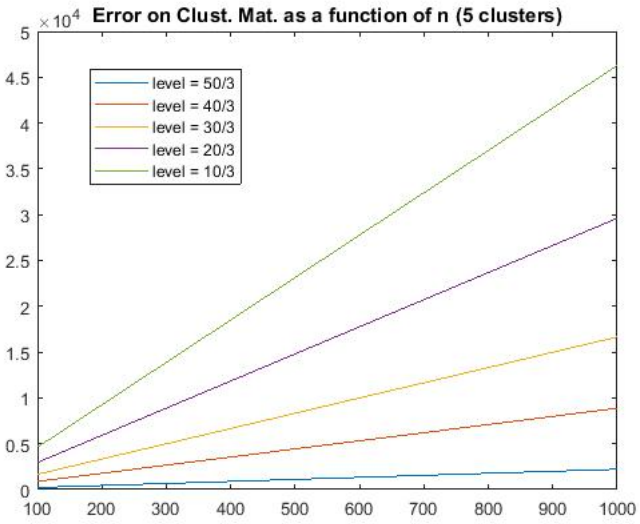


Figure 5: Estimation error $\|\bar{Z} - \hat{Z}\|_1$

conditioning standard clustering procedures. The procedure is suitable for very high dimensional data because it is based on pairwise distances only. Moreover, increasing the dimension will improve the robustness of the procedure when the Law of Large Numbers will apply along dimensions, hence forcing the affinity matrix to converge to a deterministic limit and thus making the estimator less sensitive to its low dimensional fluctuations.

Another feature of the method is that it may apply to a large number of mixtures type, even when the component's densities are not log-concave, as do a lot of embeddings as applied to data concentrated on complicated manifolds. Further studies will be performed in this exciting direction.

Future work is also needed for proving that the proposed embedding is provably efficient when combined with various clustering techniques. One of the main reason why this should be a difficult problem is that the approximation bound proved in the present paper is not so easy to leverage for controlling the perturbation of the eigenspaces of Z. More precise use of the inherent randomness of the perturbation, in the spirit of [26], might be necessary in order to go a little further in this direction.

Acknowledgements. The results presented in this paper have appeared previously as Chapter 3 of the third author's PhD thesis [9].

References

[1] Emmanuel Abbe, Afonso S Bandeira, and Georgina Hall. Exact recovery in the stochastic block model. *arXiv preprint arXiv:1405.3267*, 2014.

[2] Pranjal Awasthi, Afonso S Bandeira, Moses Charikar, Ravishankar Krishnaswamy, Soledad Villar, and Rachel Ward. Relax, no need to round: Integrality of clustering formulations. In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pages 191–200. ACM, 2015.

[3] Afonso S Bandeira. Ten lectures and forty-two open problems in the mathematics of data science. 2015.

[4] Afonso S Bandeira, Dustin G Mixon, and Benjamin Recht. Compressive classification and the rare eclipse problem. In *Compressed Sensing and its Applications*, pages 197–220. Springer, 2017.

[5] Christophe Biernacki and Stéphane Chrétien. Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with em. *Statistics & probability letters*, 61(4):373–382, 2003.

[6] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[7] Stéphane Chrétien and Alfred O Hero. On em algorithms and their proximal generalizations. *ESAIM: Probability and Statistics*, 12:308–326, 2008.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977. With discussion.

[9] Adrien Faivre. *Analyse d'image hyperspectrale*. PhD thesis, Université de Bourgogne Franche-Comté, 2018.

[10] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

[11] Olivier Guédon and Roman Vershynin. Community detection in sparse networks via grothendieck's inequality. *Probability Theory and Related Fields*, pages 1–25, 2015.

[12] Simon Heimlicher, Marc Lelarge, and Laurent Massoulié. Community detection in the labelled stochastic block model. *arXiv preprint arXiv:1209.2910*, 2012.

[13] Toby Dylan Hocking, Armand Joulin, Francis Bach, and Jean-Philippe Vert. Clusterpath an algorithm for clustering using convex fusion penalties. In *28th international conference on machine learning*, page 1, 2011.

[14] Anil K Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[15] Nathan Linial, Eran London, and Yuri Rabinovich. The geometry of graphs and some of its algorithmic applications. *Combinatorica*, 15(2):215–245, 1995.

[16] Geoffrey McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

[17] Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 2001 IEEE International Conference on Cluster Computing*, pages 529–537. IEEE, 2001.

[18] Dustin G Mixon, Soledad Villar, and Rachel Ward. Clustering subgaussian mixtures by semidefinite programming. *arXiv preprint arXiv:1602.06612*, 2016.

[19] Elchanan Mossel, Joe Neeman, and Allan Sly. Stochastic block models and reconstruction. *arXiv preprint arXiv:1202.1499*, 2012.

[20] Elchanan Mossel, Joe Neeman, and Allan Sly. Reconstruction and estimation in the planted partition model. *Probability Theory and Related Fields*, 162(3-4):431–461, 2015.

[21] Elchanan Mossel, Joe Neeman, and Allan Sly. A proof of the block model threshold conjecture. *Combinatorica*, 38(3):665–708, 2018.

[22] Peter Radchenko and Gourab Mukherjee. Consistent clustering using an ℓ_1 fusion penalty. *arXiv preprint arXiv:1412.0753*, 2014.

[23] Kean Ming Tan, Daniela Witten, et al. Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9(2):2324–2347, 2015.

[24] Joel A Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 978–986. Society for Industrial and Applied Mathematics, 2009.

[25] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[26] Van Vu. Singular vectors under random perturbation. *Random Structures & Algorithms*, 39(4):526–538, 2011.

[27] Binhuan Wang, Yilong Zhang, Wei Sun, and Yixin Fang. Sparse convex clustering. *arXiv preprint arXiv:1601.04586*, 2016.

[28] Huahua Wang and Arindam Banerjee. Bregman alternating direction method of multipliers. In *Advances in Neural Information Processing Systems*, pages 2816–2824, 2014.

A Proofs

A.1 Proof of Theorem 2.2

The proof follows the same lines as in Guédon and Vershynin [11] and we provide the main ideas for the sake of completeness. The proof is divided into 4 steps.

Step 1: We prove that

$$\langle \bar{A}, \bar{Z} \rangle - 2K_G \|A - \bar{A}\|_{\infty \to 1} \leqslant \langle \bar{A}, \widehat{Z} \rangle \leqslant \langle \bar{A}, \bar{Z} \rangle \tag{15}$$

with K_G denoting Grothendieck's constant. The upper bound follows directly from Lemma 2.1. For the lower bound, we use the definition of $\widehat{Z}$ as a maximizer and write

$$\begin{aligned} \langle \bar{A}, \widehat{Z} \rangle &= \langle A, \widehat{Z} \rangle + \langle \bar{A} - A, \widehat{Z} \rangle \\ &\geqslant \langle A, \bar{Z} \rangle - \langle A - \bar{A}, \widehat{Z} \rangle \\ &= \langle \bar{A}, \bar{Z} \rangle + \langle A - \bar{A}, \bar{Z} \rangle - \langle A - \bar{A}, \widehat{Z} \rangle. \end{aligned}$$

Grothendieck's inequality implies that for every $Z \in \mathcal{M}_{opt}$,

$$\left| \langle A - \bar{A}, Z \rangle \right| \leqslant K_G \|A - \bar{A}\|_{\infty \to 1}.$$

See Theorem C.3 and Lemma C.4 in the Appendix. Using this, we get

$$2K_G \|A - \bar{A}\|_{\infty \to 1} \geqslant \langle \bar{A}, \bar{Z} - \widehat{Z} \rangle \tag{16}$$

as desired.

Step 2: We show that for every $Z \in \mathcal{M}_{opt}$,

$$\langle \bar{A}, \bar{Z} - Z \rangle \geqslant \frac{p - q}{2} \left\| \bar{Z} - Z \right\|_1. \tag{17}$$

This corresponds to Lemma 7.2 in [11] and shows that the expected objective function distinguishes points. Introducing the set

$$\mathrm{In} = \cup_{k=1}^{K} \mathcal{C}_k \times \mathcal{C}_k \tag{18}$$

of edges within clusters and the set

$$\mathrm{Out} = \{1, \dots, n\}^2 \backslash \mathrm{In} \tag{19}$$

of edges across clusters, we decompose the scalar product

$$\langle \bar{A}, \bar{Z} - Z \rangle = \sum_{(i,j) \in \mathrm{In}} \bar{A}_{ij} (\bar{Z}_{ij} - Z_{ij}) - \sum_{(i,j) \in \mathrm{Out}} \bar{A}_{ij} (Z_{ij} - \bar{Z}_{ij}).$$

Note that the definition of the cluster matrix (2) implies that $\bar{Z}_{ij} - Z_{ij} \geqslant 0$ if $(i,j) \in \mathrm{In}$ and $\bar{Z}_{ij} - Z_{ij} \leqslant 0$ if $(i,j) \in \mathrm{Out}$. This together with condition (8) implies

$$\langle \bar{A}, \bar{Z} - Z \rangle \geqslant p \sum_{(i,j) \in \mathrm{In}} (\bar{Z} - Z)_{ij} - q \sum_{(i,j) \in \mathrm{Out}} (Z - \bar{Z})_{ij}.$$

Introduce $S_{\mathrm{In}} = \sum_{(i,j)\in \mathrm{In}} (\bar{Z} - Z)_{ij}$ and $S_{\mathrm{Out}} = \sum_{(i,j)\in \mathrm{Out}} (\bar{Z} - Z)_{ij}$. Since $\langle \bar{Z}, 1_{n\times n} \rangle = \langle Z, 1_{n\times n} \rangle = \lambda_0$, we have $S_{\mathrm{In}} - S_{\mathrm{Out}} = 0$. On the other hand $S_{\mathrm{In}} + S_{\mathrm{Out}} = \left\| \bar{Z} - Z \right\|_1$. From these computations, we easily obtain the lower bound

$$\langle \bar{A}, \bar{Z} - Z \rangle \geqslant \frac{p-q}{2} \left\| \bar{Z} - Z \right\|_1 . \tag{20}$$

Step 3: Combining (16) and (20), we obtain

$$\left\| \bar{Z} - \widehat{Z} \right\|_1 \leqslant \frac{4K_G}{p-q} \, \|A - \bar{A}\|_{\infty\to 1}. \tag{21}$$

Step 4: From (21) we get

$$\mathrm{P}\left(\left\| Z - \bar{Z} \right\|_1 > t \, n^2 \right) \leqslant \mathrm{P}\left(\left\| A - \bar{A} \right\|_{\infty\to 1} > t \, \frac{p-q}{4K_G} \, n^2 \right).$$

and follows then directly from Proposition 2.4.

Step 5: For every matrix $H \in \mathbb{R}^{n\times n}$, we have

$$\|H\|_1 \geqslant \|H\|_{\infty\to 1} . \tag{22}$$

From Proposition 5.2 in [24], there exists a subset $\tau \subset \{1, \ldots, n\}$ such that $|\tau| \geqslant \frac{n}{2}$ and

$$\|H_{\tau\times\tau}\|_1 \leqslant \frac{2K_G}{n} \, \|H\|_{\infty\to 1}.$$

Therefore, taking $H = \bar{Z} - \widehat{Z}$, we get

$$\left\| \left(\bar{Z} - \widehat{Z} \right)_{\tau\times\tau} \right\|_1 \leqslant \frac{2K_G}{n} \left\| \left(\bar{Z} - \widehat{Z} \right) \right\|_{\infty\to 1} \tag{23}$$

and Equation (22) entails

$$\left\| \left(\bar{Z} - \widehat{Z} \right)_{\tau\times\tau} \right\|_1 \leqslant \frac{2K_G}{n} \left\| \left(\bar{Z} - \widehat{Z} \right) \right\|_1 . \tag{24}$$

Combining this last equation with (21), we obtain

$$\left\| \left(\bar{Z} - \widehat{Z} \right)_{\tau\times\tau} \right\|_1 \leqslant \frac{1}{n} \frac{8K_G^2}{p-q} \, \|A - \bar{A}\|_{\infty\to 1}.$$

We thus may deduce that

$$\begin{aligned} &\mathrm{P}\left(\left\| \left(\bar{Z} - \widehat{Z} \right)_{\tau\times\tau} \right\|_1 > t \, n \right) \\ &\qquad \leqslant \mathrm{P}\left(\left\| A - \bar{A} \right\|_{\infty\to 1} > t \, \frac{p-q}{4K_G} \, n^2 \right). \end{aligned}$$

and (12) follows then directly from Proposition 2.4.

A.2 Proof of Proposition 2.4

The concentration of the affinity matrix A around its mean $\bar{A}$ follows from concentration inequalities for Lipschitz function of independent standard Gaussian variables. From definition (13)

$$\|A - \bar{A}\|_{\infty \to 1} = \max_{u,v \in \{-1,1\}^n} F_{uv} \tag{25}$$

$$F_{uv} = \sum_{i,j=1}^{n} u_i v_j (A_{i,j} - \bar{A}_{i,j}). \tag{26}$$

We introduce the standardized observations: if x_i is in cluster $\mathcal{C}_{k_i}$, i.e. $x_i \sim \mathcal{N}(\mu_{k_i}, \Sigma_{k_i})$, then $y_i = \Sigma_{k_i}^{-1/2}(x_i - \mu_{k_i})$, $1 \leqslant i \leqslant n$ are independent identically distributed random variables with standard Gaussian distribution. In view of definition (25), the random variables F_{uv} can be expressed in terms of the standardized observations

$$F_{uv}(y_1, \ldots, y_n) =$$
$$2\sum_{1 \leqslant i < j \leqslant n} u_i v_j \left[f\left(\left\| \Sigma_{k_j}^{1/2} y_j - \Sigma_{k_i}^{1/2} y_i + \mu_{k_j} - \mu_{k_i} \right\|_2 \right) - \bar{A}_{i,j}) \right].$$

We prove next that the function $F_{uv} : \mathbb{R}^{p \times n} \to \mathbb{R}$ is L-Lipschitz with $L = 2\ell\sigma n^{3/2}$. Indeed, for $(y_1, \ldots, y_n), (y'_1, \ldots, y'_n) \in \mathbb{R}^{p \times n}$, we have

$$\begin{aligned}
& \left| F_{uv}(y_1, \ldots, y_n) - F_{uv}(y'_1, \ldots, y'_n) \right| \\
\leqslant\ & \ell \sum_{1 \leqslant i \neq j \leqslant n} \|x_i - x'_i\|_2 + \|x_j - x'_j\|_2 \\
=\ & 2(n-1)\ell \sum_{i=1}^{n} \|\Sigma_{k_i}^{1/2}(y_i - y'_i)\|_2 \\
\leqslant\ & 2n\ell \sum_{i=1}^{n} \rho(\Sigma_{k_i})^{1/2} \|y_i - y'_i\|_2 \\
\leqslant\ & 2\ell\sigma n^{3/2} \|(y_1, \ldots, y_n) - (y'_1, \ldots, y'_n)\|_2.
\end{aligned}$$

In the first inequality, we use the fact that f is ℓ-Lipschitz. The second inequality relies on the fact that all the eigenvalues of $\Sigma_{k_i}^{1/2}$ are smaller that $\rho(\Sigma_{k_i})$. The last inequality relies on Cauchy-Schwarz inequality and on the definition $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} \max_{1 \leqslant k \leqslant K} \rho(\Sigma_{k_i})$.

Thanks to this Lipschitz property, the Tsirelson-Ibragimov-Sudakov inequality implies

$$\mathrm{E}\left[\exp(\theta F_{uv})\right] \leqslant \exp\left(L^2\theta^2/2\right) \quad \text{for all} \quad \theta \in \mathbb{R}$$

and we deduce that

$$\begin{aligned}
\mathrm{E}\left[\|A - \bar{A}\|_{\infty \to 1}\right] &= \mathrm{E}\left[\max_{u,v \in \{-1,1\}^n} F_{uv}\right] \\
&\leqslant \sqrt{2L^2 \log 2^n} = 2\sqrt{2\log 2}\,\ell\sigma n^2.
\end{aligned}$$

On the other hand, the function $\max_{u,v\in\{-1,1\}^n} F_{uv}$ is also L-Lipschitz and we obtain that

$$\begin{aligned}&\mathrm{P}\left(|\|A-\bar{A}\|_{\infty\to 1}-\mathrm{E}\|A-\bar{A}\|_{\infty\to 1}|>t\right)\\&\mathrm{P}\left(|\max_{u,v\in\{-1,1\}^n} F_{uv}-\mathrm{E}\max_{u,v\in\{-1,1\}^n} F_{uv}|>t\right)\\&\qquad\leqslant 2\exp\left(-\tfrac{t^2}{8L^2}\right).\end{aligned}$$

Combining these different estimates, we obtain for $t>2\sqrt{2\log 2}\ell\sigma$,

$$\begin{aligned}&\mathrm{P}(\|A-\bar{A}\|_{\infty\to 1}>tn^2)\\\leqslant\ &\mathrm{P}\Big(\big|\|A-\bar{A}\|_{\infty\to 1}-\mathrm{E}\,\|A-\bar{A}\|_{\infty\to 1}\big|\\&\qquad>(t-2\sqrt{2\log 2}\ell\sigma)n^2)\Big)\end{aligned}$$

and thus,

$$\begin{aligned}&\mathrm{P}(\|A-\bar{A}\|_{\infty\to 1}>tn^2)\\&\quad\leqslant 2\exp\left(-\frac{\left(t-2\sqrt{2\log 2}\ell\sigma\right)^2}{32\ell^2\sigma^2}n\right).\end{aligned}$$

A.3 Proof of Proposition 2.5

The proof mainly relies on the following lemma.

Lemma A.1. *Let $X\sim\mathcal{N}(\mu,\Sigma)$. If $\mu=0$, we have*

$$\mathrm{E}\left[e^{t\|X\|^2}\right]=\prod_{d=1}^{p}\left(1-2t\sigma_d^2\right)^{-1/2},\quad t\leqslant 0,$$

with $\sigma_1^2,\dots,\sigma_p^2$ the eigenvalues of Σ. More generally, for $\mu\neq 0$,

$$\mathrm{E}\left[e^{t\|X\|^2}\right]=\prod_{d=1}^{p}\exp\left(\frac{\langle\mu,v_d\rangle^2 t}{1-2t\sigma_d^2}\right)\left(1-2t\sigma_d^2\right)^{-1/2}$$

with $\sigma_1^2,\dots,\sigma_p^2$ the eigenvalues of Σ and $v_1,\dots,v_p$ the associated eigenvectors.

The proof of the proposition then follows from the fact that X_i-X_j is a Gaussian random vector with mean $\mu_{k_i}-\mu_{k_j}$ and variance $\Sigma_{k_i}+\Sigma_{k_i}$ so that the distribution of $\|X_i-X_j\|_2^2$ is related to the noncentral χ^2 distribution with p degrees of freedom. The quantity $\mathbb{E}[\exp(-\|X_i-X_j\|_2^2/h_0)]$ corresponds to the Laplace transform of the noncentral χ^2 distribution explicited in Lemma A.1.

B Solving the SDP

The problem of solving the Semi-Definite Programming (SDP) problem given by (5) and (6), despite polynomial time solvable, can be hard to tackle in practice for large datasets. Indeed, standard packages based on interior point methods do not scale beyond medium size problems of dimension on the order 500. The main difficulty resides in having to joinlty deal with the positive semi-definiteness constraint and the componentwise non-negativity constraint. In this section, we propose a new and scalable approach to this problem, based on a linearised version of the Alternating Direction of Multipliers Method (ADMM).

B.1 A Bregman ADMM

The main idea behind the ADMM approach to solving problem (5) is that it can be augmented by including a matrix variable W which will account for the positive-semi-definiteness and the constraint $\mathrm{diag}(W) = 1_n$, while the non-negativity and 'summing-to-λ' constraints can be enforced on Z, i.e.

$$\max_{W \geq 0,\ Z \geqslant 0} \langle A, W \rangle \tag{27}$$

subject to

$$\begin{aligned} \langle Z, 1_{n \times n} \rangle &= \lambda, \\ \mathrm{diag}(W) &= 1_n, \\ W &= Z. \end{aligned} \tag{28}$$

The Lagrange function associated with this problem is

$$L(W, Z, \Lambda) = \langle A, W \rangle - \langle \Lambda, W - Z \rangle. \tag{29}$$

Given a weight $\pi > 0$, the augmented Lagrangian function is given by

$$L_\pi(W, Z, \Lambda) = \langle A, W \rangle - \langle \Lambda, W - Z \rangle - \pi \|W - Z\|_F^2. \tag{30}$$

Using the notation D_{KL} for the Kullback-Leibler divergence, the standard Bregman ADMM [28] then works as follows

$$\begin{aligned} W^{(k+1)} &= \mathrm{argmax}_{W:\ W \geq 0,\ \mathrm{diag}(W) = 1_n}\ L_\pi(W, Z^{(k)}, \Lambda^{(k)}), \\ Z^{(k+1)} &= \mathrm{argmax}_{Z:\ \langle Z, 1_{n \times n} \rangle = \lambda}\ L_\pi(W^{(k)}, Z, \Lambda^{(k)}) - \beta^{-1}\ D_{KL}(Z, Z^{(k)}), \\ \Lambda^{(k+1)} &= \Lambda^{(k)} - W^{k+1} + Z^{(k+1)} \end{aligned} \tag{31}$$

where β is a penalisation weight.

Unfortunately, these iterations cannot be easily computed, due to the quadratic penalisation term associated with the augmented Lagrangian function. Moreover, the previous scheme needs to be accelerated for practical implementability.

B.2 Linearisation and acceleration using projection

In order to obtain easy to compute iterations, one possible approach is to linearise the quadratic terms.

B.2.1 Our approach: the linearised Bregman ADMM

One easy way to go about solving this problem is to linearise this quadratic term as

$$\|W - Z^{(k)}\|_F^2 = 2\left\langle W - Z^{(k)}, W^{(k)} - Z^{(k)}\right\rangle + o\left(\|X - Z^{(k)}\|_F\right) \tag{32}$$

when minimizing with respect to the variable X and

$$\|Z - W^{(k)}\|_F^2 = 2\left\langle Z - W^{(k)}, Z^{(k)} - W^{(k)}\right\rangle + o\left(\|Z - W^{(k)}\|_F\right) \tag{33}$$

when minimizing with respect to the variable Z. Let us define the linearised augmented Lagrangian function

$$L_\pi^{lin}\left(W, Z, \Lambda, W^{(k)}, Z^{(k)}\right) = \langle A, W\rangle - \langle \Lambda, W - Z\rangle - 2\pi\left\langle W - Z, W^{(k)} - Z^{(k)}\right\rangle.$$

The linearised versions of iterations (31) are obtained after disregarding the little 'o' terms and then read

$$\begin{aligned}
W^{(k+1)} &= \operatorname{argmax}_{W:\ W\geq 0,\ \operatorname{diag}(W)=1_n} \quad L_\pi^{lin}\left(W, Z^{(k)}, \Lambda, W^{(k)}, Z^{(k)}\right), \\
Z^{(k+1)} &= \operatorname{argmax}_{Z:\ \langle Z, 1_{n\times n}\rangle = \lambda} \quad L_\pi^{lin}\left(W^{(k)}, Z, \Lambda, W^{(k)}, Z^{(k)}\right), \\
\Lambda^{(k+1)} &= \Lambda^{(k)} - \mu_k\left(W^{k+1} + Z^{(k+1)}\right),
\end{aligned}$$

where μ_k is a stepsize used to stabilise the scheme. The choice of $\mu_k = C/k$ for a constant C was observed to be the most efficient in practice.

B.2.2 Explicit expressions for the iterates

Each step in these linearised iterations has an explicit closed form expression, which is given in the following lemma.

Lemma B.1. *We have that*

- *The W-iteration is given by*

$$W^{(k+1)} = n\ V_{\max} D_{\max} V_{\max}^t \tag{34}$$

where $V_{\max}$ is a matrix whose columns are eigen-vector associated with the maximum eigenvalue of $A - \Lambda + 2\pi(Z^{(k)} - W^{(k)})$ and $D_{\max}$ is a diagonal matrix with non-negative components summing to one.

- *The Z-iteration is given by*

$$\begin{aligned} \tilde{Z}^{(k+1)} &= Z^{(k)} \odot \exp\Big(\beta(\Lambda + 2\pi\ (W^{(k)} - Z^{(k)})\Big) \\ Z^{(k+1)} &= \frac{\lambda}{\sum_{i,i'=1}^{n} \tilde{Z}^{(k)}} \quad \tilde{Z}^{(k)}. \end{aligned} \tag{35}$$

Proof. The W-step involves solving the following problem

$$\tilde{W}^{(k+1)} = \mathrm{argmax}_{X \geq 0,\ \mathrm{diag}(X)=1_n} \langle A - \Lambda + 2\pi\ (Z^{(k)} - W^{(k)}), W \rangle \tag{36}$$

which is equivalent to finding the eigenvector associated with the largest eigenvalue of $A - \Lambda + 2\pi(Z^{(k)} - W^{(k)})$. The computation of the Z-step is classical in the online optimisation literature. □

B.3 Constraining Λ

In our experiments, we also enforced the additional constraint

$$A - \Lambda^{(k)} + 2\pi(Z^{(k)} - W^{(k)}) \geqslant 0 \tag{37}$$

at every step k. Based on this constraint, one easily gets that the largest eigenvectors are non-negative by the Peron-Frobenius theorem, and therefore, non-negativity of $W^{(k+1)}$ is guaranteed. Moreover, when the multiplicity of the largest eigenvalue of $M = A - \Lambda^{(l)} + 2\pi(Z^{(k)} - W^{(k)})$ is larger than one, the graph corresponding to the weighted adjacency matrix M is disconnected and the associated eigenvectors have disjoint supports, which characterises the presence of several clusters. In our experiments, enforcing the constraint (37) never appeared to preclude convergence of $\|X^{(k)} - Z^{(k)}\|_F$ to zero. In other words, (37) was always observed to be redundant.

B.4 Choosing λ

Our approach is to simply choose the value of λ than minimises the distance between the estimated cluster matrix and the observed affinity matrix. A method based on statistical model selection will be studied in a follow up paper.

C Recalls on standard results

C.1 Concentration inequalities

The following inequality is a particular case of the Log-Sobolev concentration inequality, see Theorems 5.5 and 5.6. in [6].

Theorem C.1 (Gaussian concentration inequality)**.** *Let $Y_1, \dots, Y_n$ be independent Gaussian random vectors on $\mathbb{R}^p$ with mean 0 and variance I_p. Assume that $F : \mathbb{R}^{n\times p} \to \mathbb{R}$ is Lipschitz with constant L, i.e.*

$$|F(y') - F(y)| \leqslant L\|y' - y\|_2 \quad \textit{for all } y, y' \in \mathbb{R}^{n\times p}.$$

Then the random variable $F = F(Y_1, \dots, Y_n)$ satisfies

$$\mathrm{E}[\exp(\theta\,(F - \mathrm{E}F))] \leqslant \exp(L^2\theta^2/2) \quad \textit{for all } \theta \in \mathbb{R}$$

and also

$$\mathrm{P}(|F - \mathrm{E}F| > t) \leqslant 2\exp\left(-t^2/(8L^2)\right) \quad \textit{for all } t > 0.$$

The next theorem provides useful results for the expected maxima of (non necessarily independent) subgaussian random variables.

Theorem C.2. *Let $Z_1, \cdots, Z_N$ be real valued sub-Gaussian random variables with variance factor ν, i.e. satisfying*

$$\mathrm{E}[\exp(\theta Z_i)] \leqslant \exp(\nu\theta^2/2) \quad \textit{for all } \theta \in \mathbb{R}.$$

Then

$$\mathrm{E}\left[\max_{i=1,\cdots,N} Z_i\right] \leqslant \sqrt{2\nu \log N}.$$

C.2 The Grothendieck inequality

In this paper, we use the following matrix version of Grothendieck inequality. We denote by $\mathcal{M}_G$ the set of matrices $Z = XY^T$ with $X, Y \in \mathbb{R}^{n\times n}$ having all raws in the unit Euclidean ball, i.e.

$$\forall i \in \{1, \dots, n\}, \quad \sum_{j=1}^{n} X_{ij}^2 \leqslant 1 \quad \text{and} \quad \sum_{j=1}^{n} Y_{ij}^2 \leqslant 1$$

Theorem C.3 (Grothendieck inequality)**.** *There exists an universal constant K_G such that every matrix $B \in \mathbb{R}^{n\times n}$ satisfies*

$$\max_{Z\in\mathcal{M}_G} |\langle B, Z\rangle| \leqslant K_G\|B\|_{\infty\to 1}$$

where the $\ell^\infty \to \ell^1$ norm of B is defined by (13).

It is also useful to note the following properties of $\mathcal{M}_G$, see Lemma 3.3 in [11].

Lemma C.4. *Every matrix $Z \in \mathbb{R}^{n\times n}$ such that $Z \succeq 0$ and $\mathrm{diag}(Z) \leqslant 1_n$ satisfies $Z \in \mathcal{M}_G$.*