A SemiDefinite Programming approach to Gaussian Mixture based clustering

Adrien Faivre [1,2] & Stéphane Chrétien [3] & Clément Dombry [1]

[1] *Laboratoire de Mathématiques de Besançon, UMR CNRS 6623, Université de Bourgogne Franche-Comté, 16 route de Gray, 25030 Besançon cedex, France. Email: clement.dombry@univ-fcomte.fr*
[2] *Digital Surf, 16 Rue Lavoisier, 25000 Besançon, France. Email: afaivre@digitalsurf.fr*
[3] *National Physical Laboratory, Mathematics and Modelling, Hampton Road, Teddington, TW11 OLW, UK. Email: stephane.chretien@npl.co.uk*

Résumé. Certains problèmes de classification présentent des relaxations continues obtenues par optimisation semi-définie positive et reconnues pour leur efficacité envers ces problèmes autrement insolubles. Dans un article récent, Guédon et Vershynin décrivent un nouvel angle d'attaque pour un problème voisin de détection de communautés, basé là encore sur un programme SDP. L'analyse de leur technique, reposant sur une inégalité de Grothendieck, montre à quel point elle gère efficacement le problème. Dans cet article, nous transposons cette méthode à la classification de données issues d'un modèle de mélange de gaussiennes, et nous intéressons aux garanties théoriques apportées par cette adaptation.
Mots-clés. Partitionnement de données, Modèle de mélanges gaussiens, Optimisation SDP, Relaxation continue.

Abstract. Semidefinite programming relaxations are known to deal efficiently with otherwise intractable clustering problems. In a recent article, Guédon and Vershynin devise a new way to tackle community detection, a related problem, based on the optimization of a semidefinite program. Their analysis relies on a Grothendieck inequality and the strategy turns out to be very efficient. In this paper, we adapt this strategy to a gaussian mixture clustering problem and investigate what theorical guarantees it leads to.
Keywords. Data clustering, Gaussian mixture model, Semidefinite programming, Approximation algorithms.

1 Introduction

Unsupervised clustering is a key problem in modern data analysis which has to be solved efficiently. Traditional approaches to clustering are model based (e.g. Gaussian mixture models) or nonparametric. For mixture models, the algorithm of choice has long been the EM algorithm by Dempster et al. [7], see the monograph by McLachlan and Peel [11] for an overview of finite mixture models. Nonparametric algorithms such as K-means,

K-means ++ and generalizations have been used extensively in computer science; see Jain [10] for a review. The main drawback of these standard approaches is that the minimization problems underlying the various procedures are not convex. Even worse, the log-likelihood function of e.g. Gaussian mixture model exhibits degenerate behavior, see Biernacki and Chrétien [2]. As a result, one can never certify that such algorithms have converged to an interesting stationary point and the popularity of such methods seems to be based on their satisfactory average practical performance.

Recently, very interesting results have appeared for the closely related problem of community detection based on the stochastic block model, see Abbe et al. [1], Heimlicher et al. [9] and Mossel et al. [12]. In this model, a random graph is constructed by partitioning the set of vertices V into K clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$ and by setting an edge between vertices v and v' with probability $p_{kk'}$ if $v \in \mathcal{C}_k$ and $v' \in \mathcal{C}_{k'}$. All edges are independent and the probabilities of edges depends only on the clusters structure. It is assumed that this probability is larger within clusters, i.e.

$$p = \min_{1 \leqslant k \leqslant K} p_{kk} > \max_{1 \leqslant k \neq k' \leqslant K} p_{kk'} = q. \tag{1}$$

This corresponds to the intuitive notion of cluster in graph theory where clusters have a higher edge density. Guédon and Vershynin [8] proved that the problem of recovering the clusters from the random graph can be addressed via Semi-Definite Programming (SDP) with an explicit control of the error rate. Although not explicitly studied in their paper, the SDP can be solved efficiently thanks to a general theory, see Boyd and Vandenberghe [4].

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

$$x_i \sim \mathcal{N}(\mu_k, \Sigma_k) \quad \text{if } i \in \mathcal{C}_k \tag{2}$$

with $\mu_k \in \mathbb{R}^d$ the cluster mean and $\Sigma_k \in \mathbb{R}^{d \times d}$ the cluster covariance matrix. 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 $i = 1, \ldots, n$, 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. Note that our framework slightly differs from the usual setting for Gaussian unmixing where one usually assume that the data set is made of independent observations from the mixture of Gaussian distribution and the cluster size have random sizes with multinomial distributions. The parameters of the Gaussian mixtures are to be estimated and the most popular approach is based on maximum likelihood estimation via the EM algorithm [7] and its variants like CEM, see Céleux and Govaert [6]. The likelihood may behave quite badly and exhibit degenerate behavior, making optimization via EM not always reliable, see [2]. Once all the parameters of the Gaussian mixture are estimated, the probability

that an observation belongs to a cluster are computed. Maximizing these probabilities results in a partition of the space that provides the clustering.

The goal of the present paper is to propose a complete different approach where we do not estimate the Gaussian mixture parameters but rather directly try to recover the different clusters via the solution of a Semi-Definite Programming (SDP) problem. We present how the method of Guédon and Vershynin can be adapted to the problem of Gaussian clustering with a theoretical upper bound for the misclassification rate. This adaptation is non trivial because, unlike the stochastic block model, the affinity matrix associated to Gaussian clustering does not have independent entries. Thus we need to introduce concentration inequalities for Gaussian measures, see e.g. the monograph by Boucheron et al. [3].

2 Main result

Inspired by the analysis of community detection in stochastic block model by Vershynin and Guédon [8], we propose and study a Semi-definite Program associated with Gaussian clustering. Based on the data set $x_1, \dots, x_n$, we construct an 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 $f : [0, +\infty) \to [0, 1]$ an affinity function. A popular choice is the Gaussian affinity

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

and other possibilities are

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

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 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, the Semi-Definite Program reads

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

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

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

Let us provide some intuitions for motivating the SDP problem (5). Note that each $Z \in \mathcal{M}_{opt}$ has entries in $[0,1]$ with constant sum equal to λ_0. The SDP procedure will distribute the mass λ_0 and assign more mass to entries Z_{ij} corresponding to large values of the affinity $A_{ij} = f(\|x_j - x_i\|_2)$, i.e. pairs of close points x_i, x_j. This mass distribution must respect symmetry and the constraint $Z \succeq 0$. For the analysis of the procedure, the main idea is that we want the solution $\widehat{Z}$ to be an approximation of $\bar{Z}$, the cluster matrix defined by $\bar{Z}_{i,j} = 1$ if i and j are in the same cluster, and 0 otherwise. The cluster matrix has values in $\{0,1\}$ and belongs to $\mathcal{M}_{opt}$ for $\lambda_0 = \sum_{k=1}^K n_k^2$ given by the cluster sizes, and solves the alternative SDP problem

$$\text{maximize } \langle \bar{A}, Z \rangle \quad \text{subject to} \quad Z \in \mathcal{M}_{opt} \tag{7}$$

where $\bar{A} = (\mathrm{E} f(\|x_i - x_j\|_2))_{1 \leqslant i,j \leqslant n}$ denotes the expected affinity matrix.

Our main result provides a non asymptotic upper bound for the probability that $\widehat{Z}$ differs from $\bar{Z}$ in L^1 distance.

Theorem 1 *Consider the Gaussian mixture* (2)*. Assume that the affinity function f is ℓ-Lipschitz and furthermore that*

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

Then, for all $t > t_0 = 8\sqrt{2 \log 2} K_G \sigma \ell / (p-q)$,

$$\mathrm{P}\left(\left\| \widehat{Z} - \bar{Z} \right\|_1 > n^2 t \right) \leqslant 2 \exp\left(- \left(\frac{t - t_0}{c} \right)^2 n \right), \quad c = \frac{16\sqrt{2} K_G \ell \sigma}{p-q}, \tag{9}$$

where $K_G \approx 1.7$ 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.

Theorem 1 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

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

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

Corollary 1 *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.

3 A concentration inequality

The proof is an adaptation of Vershynin and Guédon [8] whose two main ingredients are the Grothendieck inequality together with a Bernstein concentration inequality. Grothendieck's inequality is used to compare the solutions $\hat{Z}$ and $\bar{Z}$ of the two SDP programs (5) and (7). It implies that for every $Z \in \mathcal{M}_{opt}$,

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

where the ℓ^∞–ℓ^1 norm of a matrix $M \in \mathbb{R}^{n\times n}$ is defined by

$$\|M\|_{\infty\to 1} = \sup_{\|u\|_\infty \leqslant 1} \|Au\|_1 = \max_{uv\in\{-1,1\}^n} \sum_{i,j=1}^{n} u_i v_j M_{i,j}. \tag{10}$$

The concentration inequality is used to estimate $\|A - \bar{A}\|_{\infty\to 1}$ which quantifies the fluctuation of the affinity matrix A around its mean $\bar{A}$. Unlike in the stochastic block model, the entries of the affinity matrix (3) are not independent and we can not use Bernstein concentration inequality. We use Gaussian concentration instead and prove the following concentration inequality.

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

$$\mathrm{P}\Big(\left\|A - \bar{A}\right\|_{\infty\to 1} > t\, n^2\Big) \leqslant 2\exp\left(-\frac{\big(t - 2\sqrt{2\log 2}\ell\sigma\big)^2}{32\ell^2\sigma^2}\, n\right). \tag{11}$$

4 Discussion

Solving SDP problems like (5) can be done efficiently, especially for low rank matrices, using a Burer–Monteiro trick [5]. Note that the rank of Z is K. Modifying problem (5) by adding a rank-K constrained makes therefore the problem efficienly solvable.

Once we have recovered a matrix $\hat{Z}$ close enough to $\bar{Z}$, we can use spectral perturbation arguments to recover the clusters. The eigendecomposition of $\bar{Z}$ consists of K eigenvalues, $\sqrt{|\mathcal{C}_1|}$, …, $\sqrt{|\mathcal{C}_K|}$, and the corresponding eigenvectors are $1/\sqrt{|\mathcal{C}_1|}\ 1_{\mathcal{C}_1}$, …, $1/\sqrt{|\mathcal{C}_K|}\ 1_{\mathcal{C}_k}$, supposing that the cluster sizes are all different, which implies that every nonzero eigenvalue have multiplicity equal to one.

References

[1] E. Abbe, A.S. Bandeira, and G. Hall, *Exact recovery in the stochastic block model*, arXiv preprint arXiv:1405.3267 (2014).

[2] C. Biernacki and S. Chrétien, *Degeneracy in the maximum likelihood estimation of univariate gaussian mixtures with em*, Statistics & probability letters **61** (2003), no. 4, 373–382.

[3] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities*, Oxford University Press, Oxford, 2013, A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[4] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[5] S. Burer and R.D.C. Monteiro, *A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization*, Mathematical Programming **95** (2003), no. 2, 329–357.

[6] G. Celeux and G. Govaert, *A classification EM algorithm for clustering and two stochastic versions*, Comput. Statist. Data Anal. **14** (1992), no. 3, 315–332.

[7] 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** (1977), no. 1, 1–38, With discussion.

[8] O. Guédon and R. Vershynin, *Community detection in sparse networks via grothendieck's inequality*, arXiv preprint arXiv:1411.4686 (2014).

[9] S. Heimlicher, M. Lelarge, and L. Massoulié, *Community detection in the labelled stochastic block model*, arXiv preprint arXiv:1209.2910 (2012).

[10] A.K. Jain, *Data clustering: 50 years beyond k-means*, Pattern recognition letters **31** (2010), no. 8, 651–666.

[11] G. McLachlan and D. Peel, *Finite mixture models*, John Wiley & Sons, 2004.

[12] E. Mossel, J. Neeman, and A. Sly, *Stochastic block models and reconstruction*, arXiv preprint arXiv:1202.1499 (2012).