UNIVERSITY OF CALIFORNIA, SAN DIEGO

Topics in Clustering: Feature Selection and Semiparametric Modeling

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Mathematics with a Specialization in Statistics

by

Xiao Pu

Committee in charge:

Professor Ery Arias-Castro, Chair
Professor Jelena Bradic
Professor Sanjoy Dasgupta
Professor Rayan Saab
Professor Lawrence Saul

2017

The dissertation of Xiao Pu is approved, and it is acceptable in quality and form for publication on microfilm:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2017

DEDICATION

To my loving family and friends.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I owe a great deal to so many that helped along the way of my PhD study at UC San Diego:

First and foremost I would like to express my sincerest gratitude and appreciation to my advisor Professor Ery Arias-Castro. He guided me to work into the topic of Clustering, a very exciting and challenging field on the cutting edge of today's statistical research. I benefit so much from his deep statistical insights, his invaluable guidance and his constant encouragement. I feel extremely lucky to have Ery as my PhD advisor.

I would also like to thank Professor Jelena Bradic for teaching me elegant statistics. I am also very grateful to Professor Sanjoy Dasgupta and Professor Lawrence Saul for their inspiring lectures in Machine Learning and Artificial Intelligence. Sincere thanks are given to Professor Rayan Saab for serving as member of my committee.

I wish to thank the Statistics group of the Mathematics department for admitting me into the PhD program four years ago, even though I had little Statistics background at that time. I thank my classmates and friends for their sincere help. I will always cherish our friendship.

Finally, I want to thank my parents and my elder brother. Without their boundless love and support, this work would not have been completed.

Chapter 4, in full, has been published in Computational Statistics and Data Analysis. Ery Arias-Castro, Xiao Pu, "A Simple Approach to Sparse Clustering", *Computational Statistics and Data Analysis*,105 (2017): 217-228. The dissertation author is the corresponding author of this material.

Chapter 5, in full, has been organized into the following paper: *Semiparametric Estimation of Symmetric Mixture Models with Monotone and Log-Concave Densities* (Xiao Pu and Ery Arias-Castro), and has been submitted for publication. The dissertation author is the primary investigator and corresponding author of this material.

Chapter 6, in full, has been organized into the following paper: *Concentration of Measure for Radial Distributions and Consequences for Statistical Modeling* (Ery

Arias-Castro and Xiao Pu), and has been submitted for publication. The dissertation author is the corresponding author of this material.

VITA

2011	B. S. in Nuclear Engineering and Technology, Shanghai Jiao Tong University, Shanghai, China
2013	M. S. in Physics, University of Minnesota - Duluth, Duluth, MN
2017	M. S. in Statistics, University of California, San Diego
2017	Ph. D. in Mathematics with a Specialization in Statistics, University of California, San Diego

PUBLICATIONS

Ery Arias-Castro, Xiao Pu, "A Simple Approach to Sparse Clustering", *Computational Statistics and Data Analysis*,105 (2017): 217-228.

Xiao Pu, Ery Arias-Castro, "Semiparametric Estimation of Symmetric Mixture Models with Monotone and Log-Concave Densities", *arXiv preprint arXiv:1702.08897*.

Ery Arias-Castro, Xiao Pu, " Concentration of Measure for Radial Distributions and Consequences for Statistical Modeling", *arXiv preprint arXiv:1607.07549*.

ABSTRACT OF THE DISSERTATION

Topics in Clustering: Feature Selection and Semiparametric Modeling

by

Xiao Pu

Doctor of Philosophy in Mathematics with a Specialization in Statistics

University of California San Diego, 2017

Professor Ery Arias-Castro, Chair

The first part of this thesis is concerned with Sparse Clustering, which assumes that a potentially large set of features are associated with clustering observations but the true underlying clusters differ only with respect to some of the features. We propose two approaches for this purpose, both of which allow us to group the observations using only a carefully-chosen subset of the features. The first approach assumes that the data are generated from Gaussian mixture models in high dimensions and the difference between mean vectors of the Gaussian components is sparse. Enlightened by the connection between sparse principal component analysis (SPCA) and sparse clustering, we adapted multiple estimation strategies from SPCA to perform sparse clustering. We provide theoretical guarantee of the aggregated estimator and develop an iterative algorithm to uncover the important feature set in sparse clustering. The second one is a hill-climbing approach, which alternates between selecting the s most important features (that correspond to the s smallest within-cluster dissimilarities) and clustering observations based on the selected feature subset. This approach has been shown to be competitive with existing methods in literature on simulated and real-world datasets.

In the second part of the thesis, we consider a semiparametric approach to clustering and develop related theory. We first consider the problem of fitting a mixture model under the assumption that the mixture components are symmetric and log-concave. We study the nonparametric maximum likelihood estimation (NPMLE) of a monotone and log-concave probability density (which we do as part of our algorithm), and derive some results in terms of existence, uniqueness and uniform consistency of the MLE. To fit the mixture model, we propose a semiparametric EM (SEM) algorithm, which can be adapted to other semiparametric mixture models. We then consider mixture modeling in high dimensions using radial (or elliptical) distributions. In the process of working on this problem, we uncovered a difficulty in estimating the densities. We found that the i.i.d. d-dimensional data points sampled from a rotationally invariant distribution F with density $f(\boldsymbol{x}) = g(\|\boldsymbol{x}\|)$, are highly concentrated on the sphere of a d-dimensional ball as $d \to \infty$. This extends the well-known behavior of the normal distribution (its concentration around the sphere of radius square-root of the dimension) to other radial densities. We establish a form of concentration of measure, and even a convergence in distribution, under additional assumptions. We draw some possible consequences for statistical modeling in high-dimensions, including a possible universality property of Gaussian Mixtures.

Chapter 1

Introduction

Clustering objects (observations, events) into similar clusters is an important practical problem in a wide variety of fields, including statistics, physics, bioinformatics, artificial intelligence, and data mining. The definition of what constitutes a cluster is not precisely defined, therefore there are so many clustering algorithms (Estivill-Castro, 2002). In general, they can be classified into two categories: hierarchical methods and partition methods. Hierarchical clustering typically starts from a proximity matrix that captures differences between the objects to be clustered and produces a nested sequence of partitions, with a single, all-inclusive cluster at the top and singleton clusters of individual objects at the bottom. In contrast, partition algorithms usually produce non-overlapping clusters having no hierarchical relationships between them. The partitioned clusters are typically represented by a central vector and objects are assigned to the nearest cluster center. The popular K-means algorithm (MacQueen, 1967) and its variants are members of this class. A statistically motivated partition method is model-based clustering, which models the data as a sample from a mixture distribution (not necessarily Gaussian), with each component corresponding to a cluster.

With the recent advent of technologies, good clustering algorithms are very much desired for analyzing high-dimensional data where the number of variables is considerably larger than the number of objects. For supervised learning, when we are in the high-dimensional setting, we often assume that only a small subset of the original features are relevant, and a carefully-chosen subset of the features will

usually lead to better performance. Feature selection has been studied extensively in the literature for regression and classification problems (Akaike, 1974; Candes and Tao, 2007; Fan and Li, 2001; Mallows, 1973; Tibshirani, 1996; Zou and Hastie, 2005), but in the context of clustering it is still at a comparatively infant stage of development. Nevertheless, it has started receiving increased attention recently, and in Section 2.1 we review some of the main proposals in the literature. In Chapter 3 and Chapter 4, we propose two different approaches for feature selection in clustering.

The approach proposed in Chapter 3 is motivated from recent development in the computation of the sparse principal component for the spiked covariance matrix proposed by Johnstone and Lu (2009). The simplest variant of the spiked covariance model assumes that $X_1, \cdots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma)$, with

$$\Sigma = \lambda\theta\theta' + I_p, \qquad \lambda \geq 0. \tag{1.1}$$

Or equivalently,

$$X_i = \sqrt{\lambda}u_i\theta + Z_i \text{ and } \Sigma = \lambda\theta\theta' + I_p, \quad 1 \leq i \leq n, \tag{1.2}$$

where θ is a fixed vector of norm 1, $u_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $Z_i \overset{\text{iid}}{\sim} \mathcal{N}(\mathbf{0}, I_p)$, and the u's and Z's are independent. Under the spiked covariance model, Johnstone and Lu (2009) proved that when p is comparable or dominates n, standard PCA is not consistent. This finding motivates the ℓ_0 sparse PCA problem defined as follows,

$$\mathcal{L}_0(\hat{\Sigma}) = \underset{\|\theta\|_2=1, \|\theta\|_0 \leq s}{\arg\max} \; \theta'\hat{\Sigma}\theta, \tag{1.3}$$

where $\hat{\Sigma}$ is the empirical covariance matrix. ℓ_0 sparse PCA seeks to find s-sparse linear combinations of the variables that explain the most variance in the data (we say that a vector is s-sparse if it has at most s nonzero entries). Notice that this problem is combinatorially difficult and NP-hard. Cai et al. (2013) and Vu and Lei (2012) study this problem and independently establish the optimal rates for the estimation of θ, as well as the principal subspace. Several efficient algorithms such as diagonal thresholding (Johnstone and Lu, 2009), covariance thresholding (Krauthgamer et al., 2015) and semidefinite relaxation (d'Aspremont et al., 2007)

have been proposed to solve this problem. In Chapter 3, we establish the connection between sparse PCA and sparse clustering, and adapt multiple estimation strategies from sparse PCA to perform sparse clustering. We provide theoretical guarantees of the aggregated estimator and develop an iterative algorithm to uncover the important feature set in sparse clustering.

The approach presented in Chapter 4, sparse alternate sum (SAS) clustering, is a hill-climbing one in nature. It alternates between selecting s important features that correspond to the s smallest within-cluster dissimilarities and clustering observations based on the selected feature subset. This method is simple and can be cooperated with any partition clustering algorithm that applies to dissimilarities (for example, K-medoids, K-means or a spectral method). We performed a number of numerical experiments, both on simulated data and on real (microarray) data to compare this approach with other sparse clustering algorithms introduced in Section 2.1. The experiments show that our SAS clustering is competitive with these methods.

While K-means and hierarchical algorithms are largely heuristic and not based on formal models, model-based clustering offers a principled alternative. It provides a framework for incorporating our knowledge about a domain and assigns the observations to clusters via the mixture model

$$g(\mathbf{x}) = \sum_{j=1}^{k} \pi_j f_j(\mathbf{x}), \quad \sum_{j=1}^{k} \pi_j = 1, \quad \mathbf{x} \in \mathbb{R}^d, \tag{1.4}$$

where the pdf's f_j model the conditional density of the data in the jth cluster, see e.g. (McLachlan and Peel, 2000). Typically one assumes a parametric formulation $f_j(\mathbf{x}) = f(\boldsymbol{\theta}, \mathbf{x})$ for the component distributions. Depending on what we know about the underlying distribution of the data, it could be Gaussian or a member of a different family. The model can be estimated via the EM algorithm. One key advantage of using a mixture model for clustering is that it provides not only an assignment of the data to the k components, but also a measure of uncertainty for the assignment of each observation via the posterior probabilities of component membership:

$$\tau_m(X_i) := \frac{\hat{\pi}_l \hat{f}_m(X_i)}{\sum_{j=1}^{k} \hat{\pi}_j \hat{f}_j(X_i)}, \tag{1.5}$$

where $\hat{\pi}_j$ and $\hat{f}_j(X_i)$ are the estimators of π_j and $f_j(X_i)$, respectively. Two disadvantages of this approach, as has been pointed out by Chang and Walther (2007), are that its success depends on the appropriateness of the assumed parametric model, and that each model requires a different implementation of the EM algorithm based on model-specific theoretical derivations. Therefore, it is desirable to have an EM-type clustering algorithm with nonparametric or semiparametric distributions. In Section 2.2, we review some recent developments in the estimation of nonparametric and semiparametric mixture models. In Chapter 5, we propose a new algorithm to fit the location-shifted semiparametric mixture model proposed by Bordes et al. (2006) and Hunter et al. (2007):

$$g(x) = \sum_{j=1}^{k} \pi_j f(x - \mu_j), \quad \sum_{j=1}^{k} \pi_j = 1, \quad x \in \mathbb{R}, \tag{1.6}$$

where $\mu_j \in \mathbb{R}$ and f is assumed symmetric (i.e., even, $f(x) = f(-x)$ for all $x \in \mathbb{R}$). We assume that the symmetric mixture components are also log-concave. To estimate the mixture model, in Section 5.1 we first study the NPMLE of a monotone and log-concave probability density, and derive some results in terms of existence, uniqueness and uniform consistency. In Section 5.2, we propose the semiparametric EM algorithm, which has the desirable monotonicity property of a true EM algorithm and can be adapted to other semiparametric mixture models. We compare this method with that of Balabdaoui and Doss (2014) and other mixture models on both simulated and real-world datasets. Our comparison shows that our method is competitive when the data are sampled from symmetric log-concave mixtures, and the loss is small compared with fitting a Gaussian Mixture Model when the data are indeed from a normal mixture.

Nonparametric approaches to fitting multivariate mixture models can quickly become difficult in high-dimensions because of the curse of dimensionality. Additional assumptions are often needed. The most popular one might well be the Naive Bayes approach, popular in classification (Lewis, 1998), which presumes that the variables are independent, or equivalently, that the density is the product of its marginals. Another possibility is to assume that the density is elliptical, a classical assumption in multivariate analysis (Anderson, 2003), meaning that f is of the form

$f(x) = |A|g(\|Ax\|)$, where A is a positive definite matrix. In Chapter 6, we elaborate the difficulty we uncovered in estimating high-dimensional radial densities for mixture models. We find that when estimating radial densities, the sufficient statistics, the norms of the observations, are highly concentrated as the dimension d becomes large. In particular, we show a form of concentration of measure, and convergence in distribution as the dimension d increases.

This thesis interpolates material from three papers by the author and Chair of the Committee, Ery Arias-Castro. Chapter 4 uses material from (Arias-Castro and Pu, 2017). Meanwhile, Chapter 5 is based on (Pu and Arias-Castro, 2017). Finally, Chapter 6 is based on (Arias-Castro and Pu, 2016). Some material from each of these papers has also been incorporated into this introductory Chapter and Chapter 2.

Chapter 2

Related works

2.1 Review: Feature Selection in Clustering

The literature on feature selection in clustering is much smaller than that in regression or classification. Nonetheless, it is substantial and we review some of the main proposals in this section. We start with several basic definitions used in sparse clustering.

2.1.1 Notation in Sparse Clustering

Consider a typical setting for clustering n items based on pairwise dissimilarities, with $\delta(i, j)$ denoting the dissimilarity between items $i, j \in [n] := \{1, \ldots, n\}$. For concreteness, we assume that $\delta(i, j) \geq 0$ and $\delta(i, i) = 0$ for all $i, j \in [n]$. In principle, if we want to delineate κ clusters, the goal is (for example) to minimize the average within-cluster dissimilarity. In detail, a clustering into κ groups may be expressed as an assignment function $C : [n] \mapsto [\kappa]$, meaning that $C(i)$ indexes the cluster that observation $i \in [n]$ is assigned to. Let $\mathcal{C}_\kappa^n$ denote the class of clusterings of n items into κ groups. For $C \in \mathcal{C}_\kappa^n$, its average within-cluster dissimilarity is defined as

$$\Delta[C] := \sum_{k \in [\kappa]} \frac{1}{|C^{-1}(k)|} \sum_{i,j \in C^{-1}(k)} \sum \delta(i, j). \tag{2.1}$$

This dissimilarity coincides with the *within-cluster sum of squares* commonly used in k-means type of clustering algorithms, with $\delta(i, j) = \|x_i - x_j\|^2$. The resulting

optimization problem for clustering is the following:

$$\text{Given } (\delta(i,j) : i,j \in [n]), \text{ minimize } \Delta[C] \text{ over } C \in \mathcal{C}_\kappa^n. \tag{2.2}$$

This problem is combinatorial and quickly becomes computationally too expensive, even for small datasets. A number of proposals have been suggested (Hastie et al., 2009), ranging from hierarchical clustering approaches to K-medoids.

Following in the footsteps of Friedman and Meulman (2004), we consider a situation where we have at our disposal not 1 but $p \geq 2$ measures of pairwise dissimilarities on the same set of items, with $\delta_a(i,j)$ denoting the a-th dissimilarity between items $i,j \in [n]$. Obviously, these measures of dissimilarity could be combined into a single measure of dissimilarity, for example,

$$\delta(i,j) = \sum_a \delta_a(i,j). \tag{2.3}$$

Our working assumption, however, is that only a few of these measures of dissimilarity are useful for clustering purposes, but we do not know which ones. This is the setting of sparse clustering, where the number of useful measures is typically small compared to the whole set of available measures.

We assume henceforth that all dissimilarity measures are equally important (for example, when we do not have any knowledge a priori on the relative importance of these measures) and that they all satisfy

$$\sum_{i,j\in[n]} \delta_a(i,j) = 1, \quad \forall a \in [p], \tag{2.4}$$

which, in practice, can be achieved via normalization, meaning,

$$\delta_a(i,j) \leftarrow \frac{\delta_a(i,j)}{\sum_{i,j} \delta_a(i,j)}. \tag{2.5}$$

This assumption is important when combining measures in the standard setting (2.3) and in the sparse setting (2.6) below.

Suppose for now that we know that at most s measures are useful among the p measures that we are given. For $S \subset [p]$, define the S-dissimilarity as

$$\delta_S(i,j) = \sum_{a\in S} \delta_a(i,j), \tag{2.6}$$

and the corresponding average within-cluster S-dissimilarity for the cluster assignment C as

$$\Delta_S[C] := \sum_{k\in[\kappa]} \frac{1}{|C^{-1}(k)|} \sum_{i,j\in C^{-1}(k)} \delta_S(i,j). \tag{2.7}$$

If the goal is to delineate κ clusters, then a natural objective is the following:

$$\begin{gathered} \text{Given } (\delta_a(i,j) : a \in [p], i,j \in [n]), \\ \text{minimize } \Delta_S[C] \text{ over } S \subset [p] \text{ of size } s \text{ and over } C \in \mathcal{C}_\kappa^n. \end{gathered} \tag{2.8}$$

In words, the goal is to find the s measures (which play the role of features in this context) that lead to the smallest optimal average within-cluster dissimilarity. The problem stated in (2.8) is at least as hard as the problem stated in (2.2), and in particular, is computationally intractable even for small item sets.

2.1.2 COSA, sparse K-means and regularized K-means

Friedman and Meulman (2004) propose clustering objects on subsets of attributes (COSA), which (in its simplified form) amounts to the following optimization problem

$$\text{minimize} \sum_{k\in[\kappa]} \alpha(|C^{-1}(k)|) \sum_{i,j\in C^{-1}(k)} \sum_{a\in[p]} (w_a \delta_a(i,j) + \lambda w_a \log w_a), \tag{2.9}$$

$$\text{over any clustering } C \text{ and any weights } w_1,\ldots,w_p \geq 0 \text{ subject to } \sum_{a\in[p]} w_a = 1. \tag{2.10}$$

Here α is some function and $\lambda \geq 0$ is a tuning parameter. When $\alpha(u) = 1/u$, the objective function can be expressed as

$$\sum_{a\in[p]} (w_a \Delta_a[C] + \lambda w_a \log w_a). \tag{2.11}$$

When $\lambda = 0$, the minimization of (2.11) over (2.10) results in any convex combination of attributes with smallest average within-cluster dissimilarity. If this smallest dissimilarity is attained by only one attribute, then all weights will concentrate on this attribute, with weights 1 for this attribute and 0 for the others. In general, $\lambda > 0$, and the term it multiplies is the negative entropy of the weights

$(w_a : a \in [p])$ seen as a distribution on $\{1, \ldots, p\}$. This penalty term encourages the weights to spread out over the attributes. Minimizing over the weights first leads to

$$\begin{gathered} \text{minimize } \Delta_{\text{cosa}}[C] := \min_{w} \sum_{a \in [p]} (w_a \Delta_a[C] + \lambda w_a \log w_a) \\ \text{over any clustering } C, \end{gathered} \tag{2.12}$$

where the minimum is over the w's satisfying (2.10). (Note that the λ needs to be tuned.) The minimization is carried out using an alternating strategy where, starting with an initialization of the weights w (say all equal, $w_a = 1/p$ for all $a \in [p]$), the procedure alternates between optimizing with respect to the clustering assignment C and optimizing with respect to the weights. (There is a closed-form expression for that derived in that paper.) The procedure stops when achieving a local minimum.

Witten and Tibshirani (2010) observe that an application of COSA rarely results in a sparse set of features, meaning that the weights are typically spread out. They propose an alternative method, which they call Sparse K-means, which, under (2.4), amounts to the following optimization problem

$$\begin{gathered} \text{maximize } \sum_{a \in [p]} w_a \left(\tfrac{1}{n} - \Delta_a[C] \right), \\ \text{over any clustering } C \text{ and any weights } w_1, \ldots, w_p \geq 0 \\ \text{with } \|w\|_2 \leq 1, \ \|w\|_1 \leq s. \end{gathered} \tag{2.13}$$

The ℓ_1 penalty on w results in sparsity for small values of the tuning parameter s, which is tuned by the gap statistic of Tibshirani et al. (2001). The ℓ_2 penalty is also important, as without it, the solution would put all the weight on only one the attribute with smallest average within-cluster dissimilarity. A similar minimization strategy is proposed, which also results in a local optimum.

As will be shown in later sections, Sparse K-means is indeed effective in practice. However, its asymptotic consistency remains unknown. Sun et al. (2012) propose Regularized K-means clustering for high-dimensional data and prove its asymptotic consistency. This method aims at minimizing a regularized *within-cluster sum of*

squares with an adaptive group lasso penalty term on the cluster centers:

$$\text{minimize } \frac{1}{n}\sum_{k\in[\kappa]}\sum_{i\in C^{-1}(k)}\|x_i-\mu_k\|^2+\sum_{a\in[p]}\lambda_a\sqrt{\mu_{1a}^2+\cdots+\mu_{\kappa a}^2}\ , \tag{2.14}$$

over any clustering C and any sets of centers $\mu_1,\mu_2,\cdots,\mu_\kappa$.

2.1.3 Some methods for the Euclidean setting

Consider points in space (denoted $x_1,\ldots,x_n$ in $\mathbb{R}^p$) that we want to cluster. A typical dissimilarity is the Euclidean metric, denoted by $\delta(i,j)=\|x_i-x_j\|^2$. Decomposing this into coordinate components, with $x_i=(x_{ia}:a\in[p])$, and letting $\delta_a(i,j)=(x_{ia}-x_{ja})^2$, we have

$$\delta(i,j)=\sum_{a\in[p]}\delta_a(i,j). \tag{2.15}$$

A normalization would lead us to consider a weighted version of these dissimilarities. But assuming that the data have been normalized to have (Euclidean) norm 1 along each coordinate, (2.4) holds and we are within the framework described above.

This Euclidean setting has drawn most of the attention. Some papers propose to perform clustering after reducing the dimensionality of the data (Ghosh and Chinnaiyan, 2002; Liu et al., 2003; Tamayo et al., 2007). However, the preprocessing step of dimensionality reduction is typically independent of the end goal of clustering, making such approaches non-competitive.

A model-based clustering approach is based on maximizing the likelihood. Under the sparsity assumption made here, the likelihood is typically penalized. Most papers assume a Gaussian mixture model. Let $f(x;\mu,\mathbf{\Sigma})$ denote the density of the normal distribution with mean μ and covariance matrix $\mathbf{\Sigma}$. The penalized negative log-likelihood (when the goal is to obtain κ clusters) is of the form

$$-\sum_{i\in[n]}\log\Big[\sum_{k\in[\kappa]}\pi_k f_k(x_i;\mu_k,\mathbf{\Sigma}_k)\Big]+p_\lambda(\Theta), \tag{2.16}$$

where Θ gathers all the parameters, meaning, the mixture weights $\pi_1,\ldots,\pi_\kappa$, the group means $\mu_1,\ldots,\mu_\kappa$, and the group covariance matrices $\mathbf{\Sigma}_1,\ldots,\mathbf{\Sigma}_\kappa$. For instance, assuming that the data have been standardized so that each feature has

sample mean 0 and variance 1, Pan and Shen (2007) use

$$p_\lambda(\Theta) = \lambda \sum_{k\in[\kappa]} \|\mu_k\|_1. \tag{2.17}$$

This may be seen as a convex relaxation of

$$p_\lambda(\Theta) = \lambda \sum_{a\in[p]} \sum_{k\in[\kappa]} \mathbb{I}\{\mu_{ka} \neq 0\} = \lambda \sum_{k\in[\kappa]} \|\mu_k\|_0. \tag{2.18}$$

Typically, this optimization will result in some coordinates set to zero and thus deemed not useful for clustering purposes. In another variant, Wang and Zhu (2008) use

$$p_\lambda(\Theta) = \lambda \sum_{a\in[p]} \max_{k\in[\kappa]} |\mu_{ka}|. \tag{2.19}$$

To shrink the difference between every pair of cluster centers for each variable a, Guo et al. (2010) use the *pairwise fusion penalty*

$$p_\lambda(\Theta) = \lambda \sum_{a\in[p]} \sum_{1\leq k\leq k'\leq \kappa} |\mu_{ka} - \mu_{k'a}|. \tag{2.20}$$

Taking into account the covariance matrices, and assuming they are diagonal, Xie et al. (2008) use

$$p_\lambda(\Theta) = \lambda_1 \sum_{k\in[\kappa]} \sum_{a\in[p]} |\mu_{ka}| + \lambda_2 \sum_{k\in[\kappa]} \sum_{a\in[p]} |\sigma_{ka}^2 - 1|. \tag{2.21}$$

The assumption that the covariance matrices are diagonal is common in high-dimensional settings and was demonstrated to be reasonable in the context of clustering (Fraley and Raftery, 2006). Note that none of these proposals make the optimization problem (2.16) convex or otherwise tractable. The methods are implemented via an EM-type approach.

Another line of research on sparse clustering is based on coordinate-wise testing for mixing. This constitutes the feature selection step. The clustering step typically amounts to applying a clustering algorithm to the resulting feature space. For example, Jin and Wang (2014) use a Kolmogorov-Smirnov test against the normal distribution, while Jin et al. (2015) use a (chi-squared) variance test. The latter is also done in (Azizyan et al., 2013) and in (Verzelen and Arias-Castro, 2014).

This last paper also studies the case where the covariance matrix is unknown and proposes an approach via moments. In a nonparametric setting, Chan and Hall (2010) use coordinate-wise mode testing.

2.2 Review: Non- and Semi-parametric Mixture Models

The finite mixture model (1.4) typically assumes a parametric formulation $f_j(\mathbf{x}) = f(\boldsymbol{\theta}, \mathbf{x})$ for the component distributions, such as a normal model, see e.g. (Fraley and Raftery, 2002). The unknown parameters in this model can be estimated by the EM algorithm, see e.g. (Dempster et al., 1977) and (McLachlan and Krishnan, 2007). One major drawback of this model is the strong parametric assumption on the component density f_j. Problems arise when the parametric model is misspecified. Another drawback is that each model requires a specific EM algorithm based on the parametric assumption. To relax the parametric assumption, nonparametric and semiparametric approaches are becoming popular. In this section, we will review various examples in the literature.

2.2.1 Nonparametric Mixture Models

In this work, the term "nonparametric" means that no assumptions are made about the form of the f_j's in (1.4), even though the weights $\boldsymbol{\pi} = \{\pi_1, \cdots, \pi_k\}$ are scalar parameters. A standard tool in nonparametric density estimation are kernel estimators $\hat{f}_h$ based on i.i.d. data $X_1, \cdots, X_n$,

$$\hat{f}_h(x) := \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} k\left(\frac{x - X_i}{h}\right), \quad x \in \mathbb{R}, \tag{2.22}$$

where $h > 0$ is the bandwidth and $k : \mathbb{R} \to \mathbb{R}$ the kernel function. The main advantage of kernel density estimators is that they are easily computable, independent from the assumptions made on f. However, the selection of a kernel and an appropriate bandwidth in order to avoid over-smoothing or under-smoothing, is the major problem in kernel density estimation. In spite of this issue, this standard tool has been successfully applied to non- and semi-parametric estimation in

multivariate mixtures (Benaglia et al., 2009; Chang and Walther, 2007; Chauveau and Hoang, 2016; Chauveau et al., 2015; Levine et al., 2011; Mallapragada et al., 2010). More details will be provided in Section 2.2.3.

Another nonparametric approach to density estimation is to assume certain shape restrictions for f, such as monotonicity, unimodality, convexity and log-concavity. The MLE of a monotone density was first studied by Grenander (1956), who found that its NPMLE is the left derivative of the concave majorant of the empirical cumulative distribution function. For the unimodality restriction, when the true model M is known a priori, unimodal density estimation boils down to monotone estimation; when M is not known, the problem becomes difficult since the likelihood can be maximized to ∞ by placing an arbitrary large mode at some fixed observation. Several methods were proposed to remedy this problem (Bickel and Fan, 1996; Meyer and Woodroofe, 2004; Wegman, 1970; Woodroofe and Sun, 1993). Convex density estimation was pioneered by Anevski (1994) and its MLE was first studied by Jongbloed (1995) and further refined by Groeneboom et al. (2001). Log-concave densities and their applications were first studied by Bagnoli and Bergstrom (1989) and their NPMLE were extensively studied in the literature (Balabdaoui, 2004; Balabdaoui et al., 2009; Cule and Samworth, 2010; Cule et al., 2010; Doss and Wellner, 2016a; Dümbgen and Rufibach, 2009; Rufibach, 2006).

Log-concave densities generalize many densities of common parametric distributions, such as Normal, Uniform, Logistic, χ^2 or Laplace. Many other distributions, for broad ranges of their parameter values are in fact log-concave, for example, Gamma (r, λ) for $r \geq 1$, Beta (a, b) for $a \geq 1$ and $b \geq 1$, generalized Pareto, and Gumbel. Log-concave densities have lots of nice properties as described by Balabdaoui et al. (2009). One of the most fruitful applications of this family of distributions has been in the area of clustering. In the literature, EM-type clustering algorithm with nonparametric component distributions was first carried out by Chang and Walther (2007) and further extended to multivariate log-concave mixtures by Cule et al. (2010). More recently, Hu et al. (2016) studied the existence and consistency of the log-concave maximum likelihood estimator (LCMLE) of finite mixture models. Besides clustering, this LCMLE has also been applied to

mixture of regression models (Hu et al., 2017).

2.2.2 Semiparametric Mixture Models

Sometimes, the essentially nonparametric density functions in (1.4) may be partially specified by scalar parameters, a case often called semiparametric. Note that the model (1.4) is not identifiable without additional constraints. To make the model identifiable, Bordes et al. (2006) and Hunter et al. (2007) propose a univariate location-shifted semiparametric mixture model:

$$g(x) = \sum_{j=1}^{k} \pi_j f(x - \mu_j), \quad \sum_{j=1}^{k} \pi_j = 1, \quad x \in \mathbb{R}, \tag{2.23}$$

where $\mu_j \in \mathbb{R}$ and f is assumed symmetric (i.e., even, $f(x) = f(-x)$ for all $x \in \mathbb{R}$). These authors show that the parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ and f are uniquely identifiable when $k = 2$ (up to label-shifting) as long as $\pi_1 \neq 1/2$. Furthermore, Hunter et al. (2007) showed that for $k = 3$, the parameters are uniquely identifiable except when $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ take values in a particular set of Lebesgue measure zero, conjecturing that a similar result holds for any k.

Although both Bordes et al. (2006) and Hunter et al. (2007) propose methods for estimating the parameters in (2.23), these methods are inefficient and not easily generalizable beyond the case $k = 2$. Bordes et al. (2006) use the so-called minimum contrast method to estimate $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$, and use a kernel density estimation (KDE) approach which involves a model selection procedure to choose the tuning parameter. Hunter et al. (2007) employ a generalized Hodges-Lehmann estimator to estimate $\boldsymbol{\mu}$ and achieve a better rate of convergence. However, their estimator for f is not guaranteed to be a density. Bordes et al. (2007) propose a stochastic EM-like estimation algorithm which does not possess the monotone property of a genuine EM algorithm.

Model (2.23) was also studied more recently by Butucea and Vandekerkhove (2014) and Balabdaoui and Doss (2014). Butucea and Vandekerkhove propose $\sqrt{n}$-consistent M-estimators based on a Fourier approach. Balabdaoui and Doss adopt the estimators for $\boldsymbol{\pi}$ and $\boldsymbol{\mu}$ from Hunter et al. (2007) and then estimate the density f via maximum likelihood assuming it is log-concave. Note however

that, combined, their estimators for $\{\boldsymbol{\pi}, \boldsymbol{\mu}, f\}$ are not obtained by maximizing the likelihood.

2.2.3 Multivariate Non-/semi-parametric Mixtures

In the multivariate situation, the common restriction placed on the components is that each joint density $f_j(\cdot)$ is equal to the product of its marginal densities. In other words, the coordinates of the X_i vector are independent, conditional on the subpopulation or component (f_1 through f_k) from which X_i is drawn. Therefore, model (1.4) becomes

$$g(\mathbf{x}) = \sum_{j=1}^{k} \pi_j \prod_{c=1}^{d} f_{jc}(\mathbf{x}_{ic}), \tag{2.24}$$

where the function f_{jc} denotes a univariate density function. Hall and Zhou (2003) introduced this model and consider it in its full generality, while Hettmansperger and Thomas (2000) consider the special case in which the density $f_{jc}(\cdot)$ does not depend on c— that is, in which the components of X_i are not only conditionally independent but identically distributed as well:

$$g(\mathbf{x}) = \sum_{j=1}^{k} \pi_j \prod_{c=1}^{d} f_j(\mathbf{x}_{ic}). \tag{2.25}$$

What distinguishes model (2.24) from model (2.25) is the assumption in the latter that $f_{j1}(\cdot) = \cdots = f_{jd}(\cdot)$ for all j.

To encompass both the special case (2.25) and the more general case(2.24) simultaneously, Benaglia et al. (2009) introduced an intermediate case:

$$g(\mathbf{x}) = \sum_{j=1}^{k} \pi_j \prod_{c=1}^{d} f_{jb_c}(\mathbf{x}_{ic}), \tag{2.26}$$

where they allow that the coordinates of X_i are conditionally independent and that there exist *blocks* of coordinates that are also identically distributed (b_c denotes the block to which the cth coordinate belongs). These blocks may all be of size 1 so that case (2.24) is still covered, or there may exist only a single block of size r, which is the case (2.25). To fit model (2.26), an empirical "EM-like" algorithm has been introduced by Benaglia et al. (2009). It eliminates the stochasticity of

the univariate algorithm from Bordes et al. (2007), but also relies on a weighted KDE step for the updates of the f_{jb_c}'s. Moreover, this algorithm lacks any sort of theoretical justification and is not a genuine EM algorithm due to the nonparametric KDE step. Levine et al. (2011) correct this shortcoming by introducing a smoothed log-likelihood function and formulating an iterative algorithm with a provable monotonicity property.

Model (2.26) has also been modified to fit multivariate semiparametric mixtures. Benaglia et al. (2009) modify the "EM-like" algorithm to fit location-scale mixture models, and Chauveau et al. (2015) extend the smoothed versions from Levine et al. (2011) to semiparametric mixture models.

Notice that the density functions (updated by the KDE procedure) in the above-mentioned models (2.24) $\sim$ (2.26) are all univariate. Recently, Chauveau and Hoang (2016) describe a new multivariate nonparametric mixture model that extends modemix-model1l (2.24) in the sense that it allows for conditionally independent *multivariate* and *nonparametric* component densities. Equivalently, they assume that each joint density f_j in model (1.4) is equal to product of B multivariate densities that will correspond to conditionally independence multivariate *blocks* in the mixture model. They let the set of coordinates $\{1, \cdots, d\}$ be partitioned into B disjoint subsets s_l, i.e. $\{1, \cdots, d\} = \bigcup_{l=1}^{B} s_l$, where $2 \leq B < d$ is the total number of such blocks. The resulting mixture model with conditionally independent multivariate component densities is

$$g(\mathbf{x}) = \sum_{j=1}^{k} \pi_j \prod_{l=1}^{B} f_{jl}(\mathbf{x}_{is_l}). \tag{2.27}$$

Notice that updating the multivariate component densities also relies on a KDE procedure.

Since the computation time of the MLE of multivariate log-concave densities becomes quickly intractable as the dimension increases, extending log-concave mixture models to higher dimensions presents a real challenge. Chang and Walther (2007) consider a multivariate extension by assuming that the univariate marginal densities of each component are log-concave, and the dependence structure within each component is modeled with a normal copula. But they only perform simu-

lations in dimension two. With the multidimensional log-concave density studied by Cule and Samworth (2010), Cule et al. (2010) simply assume that each component f_j's in model (1.4) is log-concave and successfully apply the multivariate log-concave EM algorithm to the cancer data of Street et al. (1993) with sample size 569 and obtain only 121 misclassified instances compared to 144 with the Gaussian EM algorithm.

Chapter 3

From Sparse Principal Component Analysis (PCA) to Sparse Clustering

In this chapter, we use the parallel between sparse PCA and sparse clustering suggested by Verzelen and Arias-Castro (2014), to adapt methods developed for sparse PCA to perform sparse clustering. Under the sparse Gaussian mixture models, we adapt the aggregation method[1] of Cai et al. (2013) and derive the aggregation estimator for sparse clustering[2]. By following the theoretical analysis performed in (Cai et al., 2013), we provide theoretical guarantees of our aggregated estimator in Theorem 3. We then adapt 3 other computationally-efficient methods developed in the context of sparse PCA to perform sparse clustering. We also propose an iterative algorithm (*Iterative 2-means*) to uncover the important feature set in sparse clustering. A simulation study shows that *Iterative 2-means* outperforms the other 3 methods in terms of both sparse recovery and the estimation of the difference between mean vectors of the Gaussian components. We start this chapter with literature review of sparse PCA.

[1]This method is presented in Section 3.1 and the estimator is defined by (3.10).

[2]This estimator is described in Algorithm 1 and defined by (3.16). The estimator identifies the important feature set for clustering in high-dimensional space. This constitutes the feature selection step and the clustering step typically amounts to applying a clustering algorithm to the resulting feature space.

3.1 Sparse PCA

PCA is a classical method for reducing dimension, say from a set of observations of p possibly correlated variables into a set of values of r ($r < p$) uncorrelated variables, and is frequently used to obtain a low-dimensional representation of a dataset. It operates by projecting the data onto the r directions of maximal variance, captured by eigenvectors of the $p \times p$ population covariance matrix Σ. In practice, one does not have access to the population covariance, but instead rely on the sample variance matrix

$$\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \Sigma + \Delta, \tag{3.1}$$

where $\mathbf{X}_i, \cdots, \mathbf{X}_n$ are i.i.d. with mean $\mathbf{0}$, and Δ denotes a random noise matrix. In the classical theory of PCA, the sample eigenvectors (i.e., based on $\hat{\Sigma}$) are consistent estimators of their population analogues, when p is fixed and $n \to \infty$ (Anderson, 2004; Muirhead, 2009). However, when p is comparable to or significantly larger than n, the sample covariance matrix $\hat{\Sigma}$ may be a poor estimator to the population's covariance matrix Σ (Bickel and Levina, 2008; Lam and Fan, 2009; Levina et al., 2008), and standard PCA based on $\hat{\Sigma}$ can produce inconsistent estimates of the population's principal components (Johnstone, 2001).

Consider a $r-$principal component model, in which, when reviewed as p dimensional column vectors, observations can be written as

$$\mathbf{X}_i = \sum_{m=1}^{r} \sqrt{\lambda}_m u_{im} \theta_m + Z_i, \quad i = 1, \cdots, n \tag{3.2}$$

where θ_m's $\in \mathbb{R}^p$ are the first r principal components to be estimated, $u_{im} \overset{\text{iid}}{\sim} N(0,1)$ and $Z_i \overset{\text{iid}}{\sim} N(\mathbf{0}, \mathbf{I}_p)$. The corresponding population covariance matrix is the famous "Spiked" Covariance Model first proposed by Johnstone (2001) and then generalized by Paul (2007) :

$$\Sigma = \sum_{m=1}^{r} \lambda_m \theta_m \theta_m' + I_p \tag{3.3}$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$ and θ' is the transpose of θ. The r largest eigenvalues of Σ are $\lambda_i + 1, i = 1, \cdots, r$, and the rest are all equal to 1. The spiked covariance

model with only one spike is the following,

$$\mathbf{X}_i = \sqrt{\lambda} u_i \theta + Z_i \text{ and } \Sigma = \lambda \theta\theta' + I_p, \quad 1 \le i \le n, \tag{3.4}$$

with $1+\lambda$ the largest eigenvalue, and θ the leading eigenvector. Under this model, Johnstone and Lu (2009) first established the inconsistency of the classic PCA :

Theorem 1 (Johnstone and Lu (2009))**.** *Consider model* (3.4) *in an asymptotic setting where* $n \to \infty$ *and* $p = p(n)$ *is such that* $p/n \to \gamma \in [0, \infty)$ *and* $\lambda > 0$ *fixed.* $\hat{\theta}$ *is a normed eigenvector for the top eigenvalue of the sample covariance matrix. Then almost surely,*

$$(\hat{\theta}'\theta)^2 \to \frac{(\lambda^2 - \gamma)_+}{\lambda^2 + \gamma\lambda}.$$

In particular, the limit is 1 if and only if $\gamma = 0$, meaning $p/n \to 0$, and this is the only regime where PCA is consistent. When $\lambda \le \sqrt{\gamma}$, then $\hat{\theta}'\theta \to 0$, meaning that $\hat{\theta} \perp \theta$ in the limit. To address the inconsistency drawback, a number of authors have conducted theoretical studies and developed methodologies on sparse PCA, under the assumption that the leading eigenvectors have a certain type of sparsity. For example, with ℓ_0 sparsity constriant, the ℓ_0-sparse PCA for the top principal direction is defined as

$$\mathcal{L}_0(\hat{\Sigma}) = \underset{\|\theta\|_2=1, \|\theta\|_0 \le s}{\arg\max} \theta'\hat{\Sigma}\theta, \tag{3.5}$$

where $\hat{\Sigma}$ is the empirical sample covariance matrix. Furthermore, Vu and Lei (2012) use the ℓ_q-ball constraint ($q \in [0, 1]$), to reduce the effective number of parameters in sparse PCA and facilitate interpretation.

Theoretical analysis of sparse PCA has first been attempted on estimating the leading principal eigenvector θ_1. Johnstone and Lu (2009) first show that the classical PCA performed on a selected subset of variables with the largest sample variances leads to a consistent estimator of θ_1 if the ordered coefficients of θ_1 have rapid decay. Vu and Lei (2012) study the rates of convergence of estimation with the ℓ_q-ball constraint on θ_1, and Lounici (2013) further considers the minimax rates with missing data. Subsequently, Vu and Lei (2013) and Cai et al. (2013) independently establish the minimax error rate of estimating the principal subspace

span($\mathbf{V}$), where $\mathbf{V} = [\theta_1, \cdots, \theta_r]$ is $p \times r$ with orthonormal columns. Here, we review the assumptions, the theorem and the optimal estimation strategy given by Cai et al. (2013).

Let $\mathbf{V}_{j*}$ denote the jth row of $\mathbf{V}$. The row support of $\mathbf{V}$ is defined by

$$\text{supp}(\mathbf{V}) = \{j \in [p] : \mathbf{V}_{j*} \neq \mathbf{0}\}, \tag{3.6}$$

whose cardinality is denoted by $|\text{supp}(\mathbf{V})|$. Let the collection of $p \times r$ matrices with orthonormal columns be $\mathcal{O}(p,r) = \{\mathbf{V} \in \mathbb{R}^{p\times r} : \mathbf{V}'\mathbf{V} = I_r\}$. Define the following parameter space for Σ,

$$\Theta_0(s,p,r,\lambda) = \{\Sigma = \mathbf{V}\mathbf{\Lambda}\mathbf{V}' + I_p : 0 < \lambda \leq \lambda_r \leq \cdots \leq \lambda_1 \leq \kappa\lambda, \\ \mathbf{V} \in \mathcal{O}(p,r), |\text{supp}(\mathbf{V})| \leq s\}, \tag{3.7}$$

where $\kappa > 1$ is a fixed constant and $1 \leq r \leq s \leq p$. For two sequences of positive numbers a_n and b_n, we write $a_n \gtrsim b_n$ when $a_n \geq cb_n$ for some absolute constant $c > 0$ and $a_n \lesssim b_n$ when $b_n \gtrsim a_n$. Finally, we write $a_n \asymp b_n$ when both $a_n \gtrsim b_n$ and $a_n \lesssim b_n$ hold. The optimal rates of convergence of the subspace span($\mathbf{V}$) is then established:

Theorem 2 (Cai et al. (2013))**.** *Suppose we observe data $X_1, \cdots, X_n$ as in* (3.2)*. Let $\lambda \gtrsim \sqrt{\frac{\log n}{n}}$, $s - r \gtrsim s \wedge \log \frac{ep}{s}$ and $n \gtrsim s \log \frac{ep}{s} \vee \log \lambda$. The minimax risk for estimating the principal subspace span($\mathbf{V}$) satisfies*

$$\inf_{\hat{\mathbf{V}}} \sup_{\Sigma\in\Theta_0(s,p,r,\lambda)} \mathbb{E}||\hat{\mathbf{V}}\hat{\mathbf{V}}' - \mathbf{V}\mathbf{V}'||_F^2 \asymp \frac{\lambda+1}{n\lambda^2}\left(r(s-r) + s\log\frac{ep}{s}\right) \tag{3.8}$$

as long as the right-hand side does not exceed some absolute constant. Otherwise, there exists no consistent estimator.

The exact optimal rate is achieved via the following aggregation strategy:

1. Randomly split the sample equally according to $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \end{bmatrix}$, where $\mathbf{X}_{(i)} = \mathbf{U}_{(i)}\mathbf{D}\mathbf{V}' + \mathbf{Z}_{(i)} (i = 1, 2)$ [3] are the data sub-matrices with $p-$variate observations generated by (3.2). Compute $\mathbf{S}_{(i)} = \frac{1}{n}\mathbf{X}'_{(i)}\mathbf{X}'_{(i)}$.

[3] $\mathbf{U}_{(i)}$ is the $n/2 \times r$ random effects matrix with i.i.d. $N(0,1)$ entries, $\mathbf{D} = \text{diag}(\lambda_1^{1/2}, \cdots, \lambda_r^{1/2})$ with $\lambda_1 \geq \cdots \geq \lambda_r > 0$, $\mathbf{V}$ is $p \times r$ orthonormal and $\mathbf{Z}$ has i.i.d. $N(0, \sigma^2)$ entries which are independent of $\mathbf{U}$.

2. Let $\mathcal{B}$ denote the class of all subsets of size s from $[p]$.

3. For each index subset B in $\mathcal{B}$, construct an estimator $\hat{\mathbf{V}}_B$ by taking the r leading singular vectors of $J_B\mathbf{S}_{(1)}J_B$, where J_B is the diagonal matrix given by $(J_B)_{ii} = \mathbb{1}_{\{i\in B\}}$.

4. Set
$$B^* = \arg\max_{B\in\mathcal{B}} Tr(\hat{\mathbf{V}}_B'\mathbf{S}_{(2)}\hat{\mathbf{V}}_B) \tag{3.9}$$
and define the aggregated estimator by
$$\hat{\mathbf{V}}_* = \hat{\mathbf{V}}_{B^*}. \tag{3.10}$$

Notice that even though this estimator is asymptotically optimal, it is computationally difficult to implement because it requires going over all subsets of $[p]$ of size s. To address this issue, a number of computationally efficient approaches have been proposed in the past decade. Some of these methods are based on greedy or non-convex optimization procedures (Jolliffe et al., 2003; **?**), some are based on ℓ_1-regularization (Witten et al., 2009; Zou et al., 2006), while others are realized through semidefinite relaxations (d'Aspremont et al., 2007), or thresholding on the sample covariance matrix (Johnstone and Lu, 2009; Krauthgamer et al., 2015). The latter two approaches, due to their ability to recover ℓ_0-sparse PCs, will be introduced and adapted to sparse clustering in Section 3.2.

3.2 Extending Sparse PCA to Sparse Clustering

In this section, we will extend the theoretical study and methodologies developed for sparse PCA to sparse clustering. We will first establish the connections between these two topics. Then by following the analysis from Cai et al. (2013) under the weak ℓ_q constraint, we carry out similar analysis of spiked covariance matrix in sparse clustering and provide theoretical guarantees of their aggregation strategy in sparse clustering. Moreover, we will adapt several computationally efficient approaches in sparse PCA to the estimation of sparse clustering.

3.2.1 Connection between Sparse PCA and Sparse Clustering

Consider a high-dimensional clustering problem where we have n different p-variate vectors (with $p \gg n$) from 2 classes:

$$X_i = \mu_0 + \ell_i(\mu_1 - \mu_0) + Z_i, \tag{3.11}$$

where the class labels $\ell_1, \cdots, \ell_n \overset{\text{iid}}{\sim}$ Bernoulli $(\frac{1}{2})$ and independent of $Z_1, \cdots, Z_n \overset{\text{iid}}{\sim} N(\mathbf{0}, I_p)$. For clustering problems, the labels are unknown and the main interest is to estimate them. For sparse clustering, we are specifically interested in settings where the difference in means is sparse:

$$\Delta\mu := \mu_1 - \mu_0 \text{ is } s\text{-sparse}. \tag{3.12}$$

We say that a vector is s-sparse if it has at most s nonzero entries, where s is a fixed integer smaller or equal to p. Notice that the covariance matrix of data from model (3.11),

$$\Sigma = \frac{1}{4}\Delta\mu\Delta\mu^\top + I_p, \tag{3.13}$$

is closely related to the spike covariance model (3.4), in which $\Sigma = \lambda\theta\theta' + I_p$ for $1 \leq i \leq n$. In high-dimensional clustering problems, we usually assume that the true underlying clusters differ only with respect to some of the features. Thus, high dimensional clustering problems are often intertwined with feature selection problems. Under model (3.11), the feature selection problem reduces to identifying the support of $\Delta\mu$, which coincides with the problem of estimating the leading eigenvector in sparse PCA.

3.2.2 Theoretical Guarantee of the Aggregation Method in Sparse Clustering

Notice that when $\Delta\mu$ in (3.12) is fixed, shifting μ_0 and μ_1 will not change Σ in (3.13).To simplify the analysis, we let $\mu_0 = \mathbf{0}$ in (3.11), and denote $\mathbf{X}$ the $n \times p$ data matrix generated by

$$\mathbf{X} = 2\sqrt{\lambda}\ell\theta' + \mathbf{Z}, \tag{3.14}$$

where $\theta = \frac{\Delta\mu}{2\sqrt{\lambda}}$, ℓ is the i.i.d. Bernoulli($\frac{1}{2}$) random vector of class labels and $\mathbf{Z}$ is random noise matrix with i.i.d. $N(0,1)$ entries and independent of ℓ. Adapting the aggregation methods by Cai et al. (2013), we derive the following algorithm for sparse clustering:

Algorithm 1 Aggregation Estimation for Sparse Clustering

Require: Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$

Ensure: subset $B^* \subseteq [p]$ of cardinality k_q^* and $\hat{\theta}_{B^*}$, where k_q^* is the effective dimension defined in Definition 2.

1: Randomly split the sample equally according to $\mathbf{X} = \begin{bmatrix} \mathbf{X}_{(1)} \\ \mathbf{X}_{(2)} \end{bmatrix}$, where $\mathbf{X}_{(j)} = 2\sqrt{\lambda}\ell_{(j)}\theta' + \mathbf{Z}_{(j)}, j = 1, 2$. Denote $\mathbf{S}_{(j)} = \frac{1}{n}\mathbf{X}'_{(j)}\mathbf{X}_{(j)}$.

2: let $\mathcal{B}$ denote the class of all subsets of size k_q^* from $[p]$. For each index subset $B \in \mathcal{B}$, construct an estimator $\hat{\theta}_B$ by taking the 1st normed eigenvector of $J_B\mathbf{S}_{(1)}J_B$, where J_B is the diagonal matrix given by

$$(J_B)_{ii} = \mathbb{1}_{\{i \in B\}},$$

3: Set

$$B^* = \arg\max_{B \in \mathcal{B}} \hat{\theta}'_B \mathbf{S}_{(2)} \hat{\theta}_B, \tag{3.15}$$

and define the aggregated estimator by

$$\hat{\theta}_* = \hat{\theta}_{B^*}. \tag{3.16}$$

Remark 1. This estimator is adapted from the aggregation estimation of the principal subspace of sparse PCA by Cai et al. (2013). While they showed that the aggregation estimation is minimax optimal in weak ℓ_q space, we will only give the upper bound of our estimator in Theorem 3. However, we conjecture that it is also minimax optimal in sparse clustering under model (3.14).

Before establishing the theoretical guarantee of this aggregation method, we first define *weak-ℓ_q norm of θ* and the *effective dimension k_q^* of θ*.

Definition 1. Order the absolute value of entries in θ in decreasing order as $|\theta_{(1)}| \geq |\theta_{(2)}| \geq \cdots \geq |\theta_{(p)}|$. We define the weak-$\ell_q$ norm of θ as $\|\theta\|_{q,w} \triangleq \max_{j\in[p]} j|\theta_{(j)}|^q$. When $q = 0$, it corresponds to the ℓ_0 norm.

Definition 2. The effective dimension k_q^* of θ is defined as

$$k_q^*(s, p, n, \lambda) = \lceil x_q(s, p, n, \lambda) \rceil,$$

where $\lceil a \rceil$ denotes the smallest integer no less than $a \in \mathbb{R}$, and

$$x_q(s, p, n, \lambda) \triangleq \max \left\{ 0 \leq x \leq p : x \leq s \left(\frac{nh(\lambda)}{1 + \log(ep/x)} \right)^{q/2} \right\},$$

with $h(\lambda) = \frac{\lambda^2}{\lambda+1}$.

Here we will state and prove the main theorem of this chapter. Since this theorem is an adaption of Theorem 4 of (Cai et al., 2013), some of the arguments in the theorem are the same and some of the technical steps in the proof are nearly identical as theirs. Nevertheless, details of the theorem as well as the proof are still given for the sake of completeness.

Theorem 3. *Let $q \in [0, 2)$. Let k_q^* be defined in Definition 2. Let $\hat{\theta}_*$ be the aggregated estimator defined in* (3.16)*. Assume that*

$$\lambda \geq C_0 \sqrt{\frac{\log n}{n}}, \quad nh(\lambda) \geq C_0 k_q^* \left(1 + \log \frac{ep}{k_q^*} \right)$$

and

$$n \geq C_0 (k_q^* \log \frac{ep}{k_q^*} \vee \log \lambda) \tag{3.17}$$

for some sufficiently large constant C_0. Then there exists a constant C depending only on q such that

$$\sup_{\|\theta\|_{q,w} \leq s, \|\theta\|_2 = 1} \mathbb{E} \|\hat{\theta}_* \hat{\theta}_*' - \theta\theta'\|_F^2 \leq C\Psi(k_q^*, p, n, \lambda) \wedge 2, \tag{3.18}$$

where

$$\Psi(k_q^*, p, n, \lambda) = k_q^* \left(\frac{1 + \log \frac{ep}{k_q^*}}{nh(\lambda)} \right).$$

Before we delving into technical details of the proof, we first provide 3 Lemmas.

Lemma 1. *Let ℓ be a random vector of length n with i.i.d Bernoulli ($\frac{1}{2}$) entries, then for any $\delta \in (0,1)$,*

$$\mathbf{P}(|\frac{2}{n}\ell'\ell - 1| > \delta) \leq 2e^{-\frac{n\delta^2}{6}}.$$

Proof. Let $\mu = E(\ell'\ell) = \nu n$. With Chernoff Bound:

$$\mathbf{P}(\ell'\ell > (1+\delta)\mu) \leq e^{-\frac{\delta^2\mu}{3}} \text{ and } \mathbf{P}(\ell'\ell < (1-\delta)\mu) \leq e^{-\frac{\delta^2\mu}{2}},$$

$\mathbf{P}(|\frac{2}{n}\ell'\ell - 1| > \delta) \leq e^{-\frac{n\delta^2}{6}} + e^{-\frac{n\delta^2}{4}} \leq 2e^{-\frac{n\delta^2}{6}}$. □

Lemma 2. *Let ℓ be a random vector of length n with i.i.d Bernoulli ($1/2$) entries, $\mathbf{Z}$ be an $n \times p$ matrix with i.i.d. $N(0,1)$ entries. Then for any $b > 0$,*

$$\mathbf{P}(\|\ell'\mathbf{Z}\|^2 \geq n(1+\sqrt{p}+b)^2) \leq e^{-b^2/2}. \tag{3.19}$$

Proof. Conditioned on $\|\ell\|_0 = m$, the random variables $Y_j := \ell'\mathbf{Z}_{*\mathbf{j}} \overset{\text{iid}}{\sim} N(0,m)$ for $j = 1, \cdots, p$. Denote $\Gamma := [\frac{Y_1}{\sqrt{m}}, \cdots, \frac{Y_p}{\sqrt{m}}]$, and $\|\ell'\mathbf{Z}\|^2 = m\|\Gamma\|^2$, then by Davidson-Szarek bound (Johnson and Lindenstrauss (2001), Theorem II.7),

$$\mathbf{P}(\|\Gamma\| > 1 + \sqrt{p} + b) \leq e^{-\frac{b^2}{2}}.$$

Combining all possible values of m,

$$\begin{aligned}
&\mathbf{P}(\|\ell'\mathbf{Z}\|^2 \geq \|\ell\|_0(1+\sqrt{p}+b)^2) \\
=&\sum_{m=0}^{n} \mathbf{P}(\|\ell'\mathbf{Z}\|^2 \geq m(1+\sqrt{p}+b)^2 \;\Big|\; \|\ell\|_0 = m) \cdot \mathbf{P}(\|\ell\|_0 = m) \\
\leq&\sum_{m=0}^{n} \binom{n}{m} (\frac{1}{2})^n e^{-\frac{b^2}{2}} = e^{-\frac{b^2}{2}}
\end{aligned}$$

Clearly, (3.19) holds because $n \geq \|\ell\|_0$. □

Lemma 3 (Proposition D.1, Supplement to (Ma, 2013))**.** *Let $\mathbf{Y}$ be an $n \times k$ matrix with i.i.d. $N(0,1)$ entires. For any $t > 0$,*

$$\mathbf{P}\left\{\left\|\frac{1}{n}\mathbf{Y}'\mathbf{Y} - I_k\right\| \leq 2\left(\sqrt{\frac{k}{n}} + t\right) + \left(\sqrt{\frac{k}{n}} + t\right)^2\right\} \geq 1 - 2e^{-nt^2/2}.$$

To prove Theorem 3, we will first provide an upper bound for the estimation of θ using the sample covariance matrix under the classical setting (without sparsity), which will be directly applied to bound the *oracle risk* in the proof of Theorem 3. This upper bound is shown in Theorem 4.

Theorem 4. *Let* $\mathbf{X} = 2\sqrt{\lambda}\ell\theta' + \mathbf{Z}$*, where* ℓ *is a random vector of length* n *with i.i.d Bernoulli* $(1/2)$ *entries and* $\mathbf{Z}$ *is an* $n \times p$ *matrix with i.i.d.* $N(0,1)$ *entries. Let* $n \geq C_0(1+\log\lambda)$ *and* $\lambda \geq C_0\sqrt{(\log n/n)}$ *for some sufficiently large constant* C_0*. Let* $\hat{\theta}$ *be the first leading eigenvector of the sample covariance matrix* $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$*. Then*

$$\sup_{\|\theta\|_0=p} \mathbb{E}\|\hat{\theta}\hat{\theta}' - \theta\theta\|_F^2 \lesssim \frac{p}{nh(\lambda)} \wedge 1, \text{where } h(\lambda) = \frac{\lambda^2}{1+\lambda}. \tag{3.20}$$

Proof. The bound holds trivially when $nh(\lambda) \lesssim p$, so assume that

$$nh(\lambda) \geq C_1 p, \text{ for some sufficiently large constant } C_1. \tag{3.21}$$

Expanding $\mathbf{S} = \frac{1}{n}\mathbf{X}'\mathbf{X}$, we obtain

$$\mathbf{S} = \frac{1}{n}\big(4\lambda\theta\ell'\ell\theta' + \mathbf{Z}'\mathbf{Z} + 2\sqrt{\lambda}\theta\ell'\mathbf{Z} + 2\sqrt{\lambda}\mathbf{Z}'\ell\theta'\big)$$

Define an auxiliary matrix $\mathbf{S}_0 = \frac{4}{n}\lambda\theta\ell'\ell\theta' + I_p$ and $\Sigma = \lambda\theta\theta' + I_p$. We have

$$\|\mathbf{S} - \mathbf{S}_0\| = \|\frac{1}{n}\mathbf{Z}'\mathbf{Z} - I_p\| + \frac{4\sqrt{\lambda}}{n}\|\ell'\mathbf{Z}\|. \tag{3.22}$$

Let $t = \sqrt{\frac{3}{n}log(nh(\lambda))}$. Define the event

$$\begin{aligned} E = &\left\{|\frac{2}{n}\ell'\ell - 1| \leq \sqrt{2}t\right\} \\ &\cap \left\{\|\frac{1}{n}\mathbf{Z}'\mathbf{Z} - I_p\| \leq 2\left(\sqrt{\frac{p}{n}} + t\right) + \left(\sqrt{\frac{p}{n}} + t\right)^2\right\} \\ &\cap \left\{\|\ell'\mathbf{Z}\| \leq \sqrt{n}(1 + \sqrt{p} + \sqrt{n}t)\right\}. \end{aligned}$$

By Lemma 1, Lemma 2 and Lemma 3, there exists an absolute constant C_2 such that $P(E^c) \leq \frac{C_2}{nh(\lambda)}$. Then

$$\mathbb{E}(\|\hat{\theta}\hat{\theta}' - \theta\theta\|_F^2 \mathbb{1}_{\{E^c\}}) \leq 2C_2\frac{p}{nh(\lambda)}. \tag{3.23}$$

It remains to bound $\mathbb{E}(\|\hat{\theta}\hat{\theta}' - \theta\theta\|_F^2 \mathbb{1}_{\{E\}})$. Under the assumption that $n \geq C_0(1 + \log\lambda)$ for some large C_0, conditioned on the event E,

$$\|\mathbf{S}_0 - 2\Sigma\| = \|\lambda(\frac{4}{n}\ell'\ell - 2)\theta\theta'\| \leq 2\lambda|\frac{2}{n}\ell'\ell - 1| \leq 2\sqrt{2}t\lambda \leq \epsilon\lambda \leq \lambda/4$$

for some sufficient small $\varepsilon > 0$. Weyl's theorem (Horn and Johnson (2012), Theorem 4.3.1) then implies

$$\sigma_1(\mathbf{S}_0) \geq \sigma_1(2\Sigma) - \|\mathbf{S}_0 - 2\Sigma\| \geq 2 + 2\lambda - \lambda/4 = 2 + \frac{7}{4}\lambda. \tag{3.24}$$

Conditioned on the event E, assumption (3.21) and that $nh(\lambda) \geq C_0 \log n$ lead to $\|\frac{1}{n}ZZ' - I_p\| \leq \epsilon\lambda$ and $\sqrt{\frac{1}{n}} + \sqrt{\frac{p}{n}} + t \leq \epsilon\sqrt{\lambda}$ for some sufficiently small $\epsilon > 0$. These bounds together with (3.22), lead to

$$\|\mathbf{S} - \mathbf{S}_0\| \leq \epsilon\lambda \leq \lambda/4,$$

which, in view of Weyl's theorem (Horn and Johnson (2012), Theorem 4.3.1), leads to

$$\sigma_2(\mathbf{S}) \leq \sigma_2(\mathbf{S}_0) + \|\mathbf{S} - \mathbf{S}_0\| \leq 1 + \lambda/4. \tag{3.25}$$

Combining (3.24) and (3.25), we obtain

$$\sigma_1(\mathbf{S}_0) - \sigma_2(\mathbf{S}) \geq 1 + \frac{3}{2}\lambda. \tag{3.26}$$

Let $[\theta, \theta^\perp]$ be an orthonormal matrix. With (3.26), we apply the Sin-Theta Theorem for symmetric matrices (Davis and Kahan, 1970) on $\mathbf{S}$ and $\mathbf{S}_0$, and obtain

$$\begin{aligned}\|\hat{\theta}\hat{\theta}' - \theta\theta\|_F^2 \mathbb{1}_{\{E\}} &\leq \frac{2}{(1 + \frac{3}{2}\lambda)^2} \min\left(\|(\mathbf{S} - \mathbf{S}_0)\theta\|_F^2, \|(\mathbf{S} - \mathbf{S}_0)\theta^\perp\|_F^2\right) \mathbb{1}_{\{E\}} \\ &\leq \frac{2}{(1 + \frac{3}{2}\lambda)^2} \|(\mathbf{S} - \mathbf{S}_0)\theta\|_F^2 \mathbb{1}_{\{E\}}. \end{aligned} \tag{3.27}$$

We now control $\|(\mathbf{S} - \mathbf{S}_0)\theta\|_F^2$. Since

$$(\mathbf{S} - \mathbf{S}_0)\theta = (\frac{1}{n}\mathbf{Z}'\mathbf{Z} - I_p)\theta + \frac{2}{n}\sqrt{\lambda}\theta\ell' Z\theta + \frac{2}{n}\sqrt{\lambda}\mathbf{Z}'\ell, \tag{3.28}$$

and

$$I_p = [\theta, \theta^\perp] \begin{bmatrix} \theta \\ \theta^\perp \end{bmatrix} = \theta\theta' + \theta^\perp(\theta^\perp)', \tag{3.29}$$

$(\mathbf{S}-\mathbf{S}_0)\theta$ can be written as

$$(\mathbf{S}-\mathbf{S}_0)\theta = \theta\theta'\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}-I_p\right)\theta + \frac{1}{n}\theta^{\perp}(\theta^{\perp})'\mathbf{Z}'\mathbf{Z}\theta + \frac{2}{n}\sqrt{\lambda}\theta\ell' Z\theta + \frac{2}{n}\sqrt{\lambda}\mathbf{Z}'\ell. \quad (3.30)$$

Note that $\|AB\|_F \le \|A\|\|B\|_F$. The triangle inequality thus leads to

$$\|(\mathbf{S}-\mathbf{S}_0)\theta\|_F^2 \le \left\|\theta'\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}-I_p\right)\theta\right\|_F + \frac{1}{n}\left\|(\theta^{\perp})'\mathbf{Z}'\mathbf{Z}\theta\right\|_F + \frac{4}{n}\sqrt{\lambda}\|\mathbf{Z}'\ell\|_F. \quad (3.31)$$

Using the fact that $\mathbf{Z}$ is a random matrix with i.i.d. $N(0,1)$ entries, and that $\|\theta\| = 1$, we can compute

$$\mathbb{E}\left(\left\|\theta'\big(\frac{1}{n}\mathbf{Z}'\mathbf{Z}-I_p\big)\theta\right\|_F^2\right) = \frac{2}{n}. \quad (3.32)$$

Moreover, note that for any two independent random matrices $\mathbf{A}\in\mathbb{R}^{n\times\ell_1}$ and $\mathbf{B}\in\mathbb{R}^{n\times\ell_2}$ with i.i.d. $N(0,1)$ entires,

$$\mathbb{E}\,\|\mathbf{A}'\mathbf{B}\|_F^2 = \ell_1\ell_2\,\mathbb{E}(\langle\mathbf{A}_{*1},\mathbf{B}_{*1}\rangle) = \ell_1\ell_2 n.$$

Since $\theta(\theta^{\perp})' = \mathbf{0}$, $\mathbf{Z}\theta$ and $\mathbf{Z}\theta^{\perp}$ are independent, $\mathbb{E}\left\|(\theta^{\perp})'\mathbf{Z}'\mathbf{Z}\theta\right\|_F^2 = (p-1)n$. Hence,

$$\mathbb{E}(\|(\mathbf{S}-\mathbf{S}_0)\theta\|_F^2\mathbb{1}_{\{E\}}) \le \frac{C_3}{n}(1+p+16\lambda(1+\sqrt{p}+\sqrt{nt})^2) \quad (3.33)$$

for some absolute constant C_3. Combining (3.33), (3.27) and (3.23) leads to the conclusion. □

Following the proof of Theorem 4 in Cai et al. (2013), we prove the main theorem in this chapter, Theorem 3.

Proof of Theorem 3. Before delving into the details, we give an outline of the proof as follows:

1. We find a good sparse approximation of the true leading principal eigenvector which lies in the weak-ℓ_q ball defined in Definition 1 .

2. We decompose the risk into a summation of three terms, namely the *approximation error*, *oracle risk* and *excess risk*. The *oracle risk* is upper bounded by Theorem 3 and *approximation error* will be bounded in Step 1.

3. The excess risk is controlled by a carefully concentration-of-measure analysis in Step 3, which forms the core of the proof.

Step 1: Sparse Approximation. Fix θ with $\|\theta\|_2 = 1$ and $\|\theta\|_{q,w} \leq s$. We assume that $q > 0$. Let $\mathcal{B}(k) = \{B \subset [p] : |B| = k\}$. Let $A \in \mathcal{B}$ denote the collection of indices of θ corresponding to the k largest absolute value. Put

$$\tilde{\Sigma} = J_A \Sigma J_A + J_{A^c} = \lambda J_A \theta\theta' J_A + I_p,$$

where J_A is the diagonal matrix given by $(J_A)_{ii} = \mathbb{1}_{\{i \in A\}}$. Denote the SVD of $\lambda J_A \theta\theta' J_A$ by $\tilde{\lambda}\tilde{\theta}\tilde{\theta}'$ with $\tilde{\theta} = \alpha^{-1} J_A \theta, \tilde{\lambda} = \lambda\alpha^2$, where $\alpha = \|J_A\theta\|_2$. Observe that $\|\theta\theta' - \tilde{\theta}\tilde{\theta}'\|_F^2 = 2(1-\alpha^2)$ and $\alpha^2 = \theta_{(1)}^2 + \cdots + \theta_{(k)}^2$. Notice that $\max\{\theta_{(1)}^q, 2\theta_{(2)}^q \cdots, j\theta_{(j)}^q, \cdots\} \leq s$, hence

$$\alpha^2 = 1 - \sum_{i>k} \theta_{(i)}^2 \geq 1 - \sum_{i>k} \left(\frac{s}{i}\right)^{2/q} \geq 1 - s^{2/q} \int_k^\infty x^{-2/q} dx = 1 - \frac{q}{2-q} k \left(\frac{s}{k}\right)^{2/q}.$$

Hence,

$$\|\theta\theta' - \tilde{\theta}\tilde{\theta}'\|_F^2 \leq \frac{2q}{2-q} k \left(\frac{s}{k}\right)^{\frac{2}{q}} \leq \frac{2q}{2-q} \Psi(k, p, n, \lambda). \tag{3.34}$$

The last inequlity follows from the choice of $k = k_q^*$ defined in Definition 2. Note that if $q = 0$, this step is superfluous since θ is already sparse, and we define $\tilde{\theta} = \theta$.

Step 2: Risk Decomposition. Since $\|\theta\| = \|\hat{\theta}_*\| = 1$, we have

$$\langle \Sigma, \theta\theta' - \hat{\theta}_*\hat{\theta}_*' \rangle = \langle \lambda\theta\theta', \theta\theta' - \hat{\theta}_*\hat{\theta}_*' \rangle = \lambda(1 - \mathrm{Tr}(\theta'\hat{\theta}_*\hat{\theta}_*'\theta)) = \frac{\lambda}{2}\|\hat{\theta}_*\hat{\theta}_*' - \theta\theta'\|_F^2.$$

Therefore,

$$\begin{aligned}
&\frac{\lambda}{2}\|\hat{\theta}_*\hat{\theta}_*' - \theta\theta'\|_F^2 \\
=&\langle \Sigma, \theta\theta' - \hat{\theta}_*\hat{\theta}_*' \rangle \\
=&\langle \Sigma, \theta\theta' - \tilde{\theta}\tilde{\theta}' \rangle + \langle \Sigma, \tilde{\theta}\tilde{\theta}' - \hat{\theta}_A\hat{\theta}_A' \rangle + \langle \Sigma, \hat{\theta}_A\hat{\theta}_A' - \hat{\theta}_*\hat{\theta}_*' \rangle \\
\leq& \underbrace{\frac{\lambda}{2}\|\theta\theta' - \tilde{\theta}\tilde{\theta}'\|_F^2}_{\text{approximation error}} + \underbrace{\frac{\lambda}{2}\|\tilde{\theta}\tilde{\theta}' - \hat{\theta}_A\hat{\theta}_A'\|_F^2}_{\text{oracle risk}} + \underbrace{\left\langle \Sigma - \frac{\mathbf{S}_{(2)}}{2}, \hat{\theta}_A\hat{\theta}_A' - \hat{\theta}_*\hat{\theta}_*' \right\rangle}_{\text{excess risk}},
\end{aligned} \tag{3.35}$$

where in the last inequality, the first two items are respectively equal and the inequality is due to the fact that $\hat{\theta}_*$ is the maximizer in (3.15) and hence $\langle \frac{\mathbf{S}_{(2)}}{2}, \hat{\theta}_A\hat{\theta}_A' -$

$\hat{\theta}_*\hat{\theta}'_*\rangle \le 0$.

Step 3: Excess Risk. Write

$$\mathbf{S}_{(2)} = \frac{1}{n}X'_{(2)}X_{(2)} = \frac{1}{n}\big(4\lambda\theta\ell'_{(2)}\ell_{(2)}\theta' + \mathbf{Z}'_{(2)}\mathbf{Z}_{(2)} + 2\sqrt{\lambda}\theta\ell'_{(2)}\mathbf{Z}_{(2)} + 2\sqrt{\lambda}\mathbf{Z}'_{(2)}\ell_{(2)}\theta'\big),$$

then

$$\Sigma - \frac{\mathbf{S}_{(2)}}{2} = G + H, \tag{3.36}$$

where

$$G \triangleq \lambda\big(1 - \frac{2}{n}\ell'_{(2)}\ell_{(2)}\big)\theta\theta',$$
$$H \triangleq I_p - \frac{1}{2n}\mathbf{Z}'_{(2)}\mathbf{Z}_{(2)} - \frac{\sqrt{\lambda}}{n}\theta\ell'_{(2)}\mathbf{Z}_{(2)} - \frac{\sqrt{\lambda}}{n}\mathbf{Z}'_{(2)}\ell_{(2)}\theta'.$$

We first deal with the inner product with G: write

$$\langle G, \hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_*\hat{\theta}'_*\rangle = \langle G, \hat{\theta}_A\hat{\theta}'_A - \theta\theta'\rangle - \langle G, \hat{\theta}_*\hat{\theta}'_* - \theta\theta'\rangle.$$

Note that

$$\begin{aligned}
\langle G, \theta\theta' - \hat{\theta}_A\hat{\theta}'_A\rangle =& \left\langle \lambda\left(1 - \frac{2}{n}\ell'_{(2)}\ell_{(2)}\right), \theta'(\theta\theta' - \hat{\theta}_A\hat{\theta}'_A)\theta \right\rangle \\
=& \lambda\left(1 - \frac{2}{n}\ell'_{(2)}\ell_{(2)}\right)(1 - \theta'\hat{\theta}_A\hat{\theta}'_A\theta) \\
=& \frac{\lambda}{2}\left(1 - \frac{2}{n}\ell'_{(2)}\ell_{(2)}\right)\|\theta\theta' - \hat{\theta}_A\hat{\theta}'_A\|_F^2 \\
\le& \frac{\lambda}{2}\left|\frac{2}{n}\ell'_{(2)}\ell_{(2)} - 1\right|\|\theta\theta' - \hat{\theta}_A\hat{\theta}'_A\|_F^2.
\end{aligned} \tag{3.37}$$

Similarly, we have

$$\langle G, \hat{\theta}_*\hat{\theta}'_* - \theta\theta'\rangle \le \frac{\lambda}{2}\left|\frac{2}{n}\ell'_{(2)}\ell_{(2)} - 1\right|\|\theta\theta' - \hat{\theta}_*\hat{\theta}'_*\|_F^2. \tag{3.38}$$

Next, we will control the inner product with H: recall that $A = \mathrm{supp}(\tilde{\theta})$ is fixed. We define a collection of $p \times p$ symmetric matrices indexed by $B = \mathcal{B}(k)$ as follows:

$$K_B \triangleq \|\hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_B\hat{\theta}'_B\|_F^{-1}(\hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_B\hat{\theta}'_B)$$

which has zero trace and unit Frobenius norm. Recall that $\hat{\theta}_* = \hat{\theta}_{B^*}$, then

$$\begin{aligned}
\langle H, \hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_*\hat{\theta}'_*\rangle =& \|\hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_*\hat{\theta}'_*\|_F\langle H, K_{B^*}\rangle \\
\le& \|\hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_*\hat{\theta}'_*\|_F^2 \underbrace{\max_{B\in\mathcal{B}(k)} |\langle H, K_B\rangle|}_{\triangleq T}.
\end{aligned} \tag{3.39}$$

Assembing (3.36), (3.37), (3.38) and (3.39), we can upper bound the excess risk by

$$\begin{aligned}&\left\langle \Sigma - \frac{\mathbf{S}_{(2)}}{2}, \hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_*\hat{\theta}'_* \right\rangle \\ \leq &\frac{\lambda}{2}\left|\frac{2}{n}\ell'_{(2)}\ell_{(2)} - 1\right| \left(\|\theta\theta' - \hat{\theta}_A\hat{\theta}'_A\|_F^2 + \|\theta\theta' - \hat{\theta}_*\hat{\theta}'_*\|_F^2\right) + T\|\hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_*\hat{\theta}'_*\|_F^2. \quad (3.40)\end{aligned}$$

Combining the risk decomposition (3.35) and (3.40), we have

$$\begin{aligned}&\frac{\lambda}{2}\|\hat{\theta}_*\hat{\theta}'_* - \theta\theta'\|_F^2 \\ &\leq \frac{\lambda}{2}(\|\theta\theta' - \tilde{\theta}\tilde{\theta}'\|_F^2 + \|\tilde{\theta}\tilde{\theta}' - \hat{\theta}_A\hat{\theta}'_A\|_F^2) + T\|\hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_*\hat{\theta}'_*\|_F^2 \\ &\quad + \frac{\lambda}{2}\left|\frac{2}{n}\ell'_{(2)}\ell_{(2)} - 1\right| \left(\|\theta\theta' - \hat{\theta}_A\hat{\theta}'_A\|_F^2 + \|\theta\theta' - \hat{\theta}_*\hat{\theta}'_*\|_F^2\right). \quad (3.41)\end{aligned}$$

To simply notation, denote

$$\begin{aligned}\delta = \|\hat{\theta}_*\hat{\theta}'_* - \theta\theta'\|_F, \quad \Delta = \|\theta\theta' - \tilde{\theta}\tilde{\theta}'\|_F, \\ R = \|\tilde{\theta}\tilde{\theta}' - \hat{\theta}_A\hat{\theta}'_A\|_F, \quad M = |\frac{2}{n}\ell'_{(2)}\ell_{(2)} - 1|.\end{aligned}$$

Since

$$\|\theta\theta' - \hat{\theta}_A\hat{\theta}'_A\|_F^2 = \|\theta\theta' - \tilde{\theta}\tilde{\theta}' + \tilde{\theta}\tilde{\theta}' - \hat{\theta}_A\hat{\theta}'_A\|_F^2 \leq (\Delta + R)^2 \leq 2(\Delta^2 + R^2),$$

and $\|\hat{\theta}_A\hat{\theta}'_A - \hat{\theta}_*\hat{\theta}'_*\| \leq \delta + \Delta + R$, we have

$$\frac{\lambda}{2}\delta^2 \leq \frac{\lambda}{2}(\Delta^2 + R^2) + \lambda M(\Delta^2 + R^2) + \frac{\lambda}{2}M\delta^2 + T(\delta + \Delta + R), \quad (3.42)$$

which is equivalent to

$$(\frac{\lambda}{2} - \frac{\lambda}{2}M)\delta^2 \leq T\delta + (\Delta^2 + R^2)(\frac{\lambda}{2} + \lambda M) + T(R + \Delta). \quad (3.43)$$

Introduce the event $E_1 = \{M \leq \sqrt{2}t\}$ with $t = \sqrt{\frac{3\log(c'nh(\lambda))}{n}}$, where c' is sufficiently small such that $t \leq \frac{1}{2\sqrt{2}}$. By Lemma 1,

$$\mathbf{P}(E_1^c) \leq 2e^{-\frac{n}{6}(\sqrt{2}t)^2} = \frac{2}{c'nh(\lambda)}. \quad (3.44)$$

Conditioning on the event E_1, $M \leq \frac{1}{2}$, by Lemma 2 of Cai et al. (2013), we have

$$\delta^2 \leq \frac{16T^2}{\lambda^2} + \frac{2\big((\Delta^2 + R^2)\frac{3}{4}\lambda + T(R + \Delta)\big)}{\frac{\lambda}{4}}. \quad (3.45)$$

Therefore,

$$\begin{aligned}\mathbb{E}\delta^2 &\le \frac{16\mathbb{E}T^2}{\lambda^2} + 6(\Delta^2 + \mathbb{E}R^2) + 8\frac{\mathbb{E}[T(R+\Delta)]}{\lambda} + 4\mathbf{P}(E_1^c) \\ &= \mathbb{E}\left[\big(\frac{4T}{\lambda} + R + \Delta\big)^2\right] - \mathbb{E}\left[(R+\Delta)^2\right] + 6(\Delta^2 + \mathbb{E}R^2) + 4\mathbf{P}(E_1^c) \\ &\le 3\big(\mathbb{E}(\frac{4T}{\lambda})^2 + \mathbb{E}\,R^2 + \Delta^2\big) + 5(\Delta^2 + \mathbb{E}R^2) + \frac{1}{c'nh(\lambda)} \\ &\le \frac{48\mathbb{E}T^2}{\lambda^2} + 8(\Delta^2 + \mathbb{E}R^2) + \frac{1}{c'nh(\lambda)} \end{aligned} \tag{3.46}$$

In view of the oracle upper bound (Theorem 4), we have

$$\mathbb{E}R^2 \le C_1(1 \wedge \frac{1}{c'nh(\lambda)}). \tag{3.47}$$

Also the approximation error is bounded by

$$\Delta^2 \le \frac{2q}{2-q}\Psi(k,p,n,\lambda). \tag{3.48}$$

If $q = 0$ then $\Delta = 0$. To control the right-hand side of (3.46), it boils down to upper bound $\mathbb{E}T^2$. In the sequel we shall prove that

$$\mathbb{E}T^2 \le C_2(1+\lambda_1)\frac{k}{n}\log\frac{ep}{k} \tag{3.49}$$

for some absolutely constant C_2. Plugging (3.47), (3.48) and (3.49) into (3.46), we arrive at

$$\begin{aligned}&\|\hat{\theta}_*\hat{\theta}_*' - \theta\theta'\|_F^2 \\ &\quad\le \frac{48C_2k}{nh(\lambda)}\log\frac{ep}{k} + \frac{16q}{2-q}\Psi(k,p,n,\lambda) + \frac{8C_1(k-1)}{nh(\lambda)} + \frac{1}{c'nh(\lambda)} \\ &\quad\le \big(\max\{48C_2, 8C_1 + \frac{1}{c'}\} + \frac{16q}{2-q}\big)\Psi(k,p,n,\lambda). \end{aligned} \tag{3.50}$$

To finish the proof of the theorem, it remains to establish (3.49). To this end, recall that K_B is symmetric and $\mathrm{Tr}(K_B) = 0$. By the definition of T and H, we have

$$T = \max_{B\in\mathcal{B}(k)} |\langle H, K_B\rangle| \le T_1 + 2T_2, \tag{3.51}$$

where

$$T_1 \triangleq \frac{1}{2n}\max_{B\in\mathcal{B}(k)} |\langle \mathbf{Z}_{(2)}'\mathbf{Z}_{(2)}, K_B\rangle| \tag{3.52}$$

and

$$T_2 \triangleq \frac{1}{n}\max_{B\in\mathcal{B}(k)}|\langle\sqrt{\lambda}\mathbf{Z}'_{(2)}\ell_{(2)}\theta', K_B\rangle| = \frac{1}{n}\max_{B\in\mathcal{B}(k)}|\langle\sqrt{\lambda}\theta\ell'_{(2)}\mathbf{Z}_{(2)}, K_B\rangle| \tag{3.53}$$

As has been proved by Cai et al. (2013) (Proof of (96)),

$$\mathbb{E}T_1^2 \le \frac{24k}{n}\log\frac{ep}{k} + \frac{32k^2}{n^2}\log^2\frac{ep}{k} + \frac{62}{n}. \tag{3.54}$$

We shall prove that

$$\mathbb{E}T_2^2 \le \lambda\left(\frac{C_3}{n} + \frac{C_4k}{n}\log\frac{ep}{k} + \frac{C_5k^2}{n^2}\log^2\frac{ep}{k}\right) \tag{3.55}$$

for some sufficiently large constant C_3, C_4, C_5. Assembling (3.51) with (3.52) - (3.55) and use the fact that $(a+b)^2 \le 2(a^2+b^2)$, we arrive at

$$\begin{aligned}\mathbb{E}T^2 &\le 2\mathbb{E}T_1^2 + 8\mathbb{E}T_2^2\\ &\le C(1+\lambda)\left(\frac{k}{n}\log\frac{ep}{k} + \frac{k^2}{n^2}\log^2\frac{ep}{k}\right)\\ &\le C(1+\lambda)\frac{k}{n}\log\frac{ep}{k},\end{aligned}$$

where we used $\frac{k}{n}\log\frac{ep}{k} \le 1$ implied by the assumption (3.17).

It then remains to establish (3.55). Fix $B \in \mathbb{B}(k)$. Since $\ell_{(2)} \perp\!\!\!\perp \mathbf{Z}_{(2)}$, conditioned on the realization of $\ell_{(2)}$, $\langle\sqrt{\lambda}\theta\ell'_{(2)}\mathbf{Z}_{(2)}, K_B\rangle = \langle\sqrt{\lambda}K_B\theta\ell'_{(2)}, \mathbf{Z}_{(2)}\rangle$ is distributed as $N(0, \lambda\|K_B\theta\ell'_{(2)}\|_F^2)$. Therefore

$$\langle\sqrt{\lambda}\theta\ell'_{(2)}\mathbf{Z}_{(2)}, K_B\rangle \overset{\text{(d)}}{=} \sqrt{\lambda}\|K_B\theta\ell'_{(2)}\|_F W \tag{3.56}$$

for some $W \sim N(0,1)$ independent of $\ell_{(2)}$. Using the fact that $\|AB\|_F \le \|A\|_F\|B\|$, we have

$$\|K_B\theta\ell'_{(2)}\|_F \le \|K_B\|_F\|\theta\|\|\ell_{(2)}\| = \|\ell_{(2)}\|.$$

Consequently, $\frac{1}{n}\langle\sqrt{\lambda}\theta\ell'_{(2)}\mathbf{Z}_{(2)}, K_B\rangle$ is stochastically dominated by $\frac{1}{n}\sqrt{\lambda}\|\ell_{(2)}\||W|$. By Chernoff bound in Lemma 1,

$$\mathbf{P}\left(\|\ell_{(2)}\| > \sqrt{n/2 + t\sqrt{3n/2}}\right) \le e^{-t^2} \tag{3.57}$$

for $0 < t < \sqrt{\frac{n}{6}}$. It is easy to show that

$$\mathbf{P}(|W| \ge \sqrt{2}t) \le e^{-t^2}. \tag{3.58}$$

Fact 1. if $\mathbf{P}(A > a) \le c_1$ and $\mathbf{P}(B > b) \le c_2$ with $A > 0, B > 0$ and $A \perp\!\!\!\perp B$, then $\mathbf{P}(AB > ab) < c_1 + c_2$,

Applying the above fact, we have

$$\mathbf{P}\bigg(\|\ell_{(2)}\||W| > \sqrt{n/2 + t\sqrt{3n/2}} \cdot \sqrt{2}t \bigg) < 2e^{-t^2}. \tag{3.59}$$

Let $f(t) = \sqrt{n/2 + t\sqrt{3n/2}} \cdot \sqrt{2}t$ and $N = \binom{p}{k}$. Fix B, let $\|\ell_{(2)}\||W| = |A_B|$, then

$$\begin{aligned}
\mathbb{E} \max_{B \in \mathcal{B}} A_B^2 = \mathbb{E} \max_{i \in [N]} A_i^2 &= 2 \int_0^\infty \mathbf{P}\big(\max_{i \in [N]} |A_i| \ge f \big) f df \\
&= 2 \int_0^\infty \Big[1 - \mathbf{P}\big(\max_{i \in [N]} |A_i| \le f \big) \Big] f f' dt \\
&= 2 \int_0^\infty \big[1 - \mathbf{P}^N \big(|A_1| \le f \big) \big] f f' dt \\
&\le 2 \int_0^\infty \Big[1 - (1 - 2e^{-t^2})^N \Big] f f' dt \\
&\le 2 \int_0^\infty \big(2Ne^{-t^2} \wedge 1 \big) f f' dt
\end{aligned}$$

Since $f^2 = nt^2 + \sqrt{6n}t^3$, $2ff' = 2nt + 3\sqrt{6n}t^2$. Therefore,

$$\begin{aligned}
\mathbb{E} \max_{i \in [N]} A_i^2 &\le \int_0^\infty \big(2Ne^{-t^2} \wedge 1 \big)(2nt + 3\sqrt{6n}t^2) dt \\
&= \int_0^{\sqrt{\log 2N}} (2nt + 3\sqrt{6n}t^2) dt + \int_{\sqrt{\log 2N}}^\infty 2Ne^{-t^2}(2nt + 3\sqrt{6n}t^2) dt \\
&\le n \log 2N + \sqrt{6n}(\log 2N)^{3/2} + \int_{\sqrt{\log 2N}}^\infty 2Ne^{-t^2}(2nt + 3\sqrt{6n}t^3) dt \\
&= n \log 2N + \sqrt{6n}(\log 2N)^{3/2} + \frac{3\sqrt{6n}}{2} + \frac{3\sqrt{6n}}{2} \log 2N + n.
\end{aligned}$$

Using the fact that $\log N = \log \binom{p}{k} \le k \log \frac{ep}{k}$ and the assumption (3.17),

$$\begin{aligned}
\mathbb{E}T_2^2 &\le \frac{\lambda}{n^2} \mathbb{E} \max_{i \in [N]} A_i^2 \\
&\le \lambda \bigg(\frac{\log 2N}{n} + \frac{6n}{n^2} (\log 2N)^{3/2} + \frac{3\sqrt{6n}}{2n^2} + \frac{3\sqrt{6n}}{2n^2} \log 2N + \frac{1}{n} \bigg) \\
&\le \lambda \bigg(\frac{C_3}{n} + \frac{C_4 k}{n} \log \frac{ep}{k} + \frac{C_5 k^2}{n^2} \log^2 \frac{ep}{k} \bigg)
\end{aligned}$$

□

3.2.3 Computationally Efficient Methods

Notice that even though the aggregation estimator has theoretical guarantees, it is computationally difficult to implement because it requires going over all subsets of $[p]$ of size k_q^*. In this section, we adapt 3 estimators in sparse PCA, which are computationally efficient and can be adapted to sparse clustering.

Algorithm 2 SDP-Estimator (d'Aspremont et al., 2007)

Require: Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, sparsity level s

Ensure: vector $\hat{\theta} \in \mathbb{R}^p$

1: Let $\hat{\Sigma} = \frac{1}{n}(\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{X} - \bar{\mathbf{X}})$

2: Compute a solution $\Theta \in \mathbb{R}^{p \times p}$ of the semidefinite programming:

$$\arg\max_{\Theta} \left\{ \langle \hat{\Sigma}, \Theta \rangle : \Theta \in \mathcal{S}_+^p, tr(\Theta) = 1, \sum_{i,j} |X_{ij}| \leq s \right\},$$

where $\mathcal{S}_+^p = \{\Theta \in \mathbb{R}^{p \times p} : \Theta = \Theta^T, \Theta \succeq 0\}$ is the cone of the symmetric positive semidefinite matrices.

3: Let $\hat{\mu}$ be the top (unit-length) eigenvector of Θ. Let $S \subset [p]$ be the set of coordinates corresponding to the s largest absolute values in $\hat{\mu}$ and the resulting SDP estimator is the unit p-dimensional vector $\hat{\theta}$ with

$$\hat{\theta}_j = \begin{cases} \frac{\hat{\mu}_j}{\|\hat{\mu}_S\|} & \text{if} \quad j \in S; \\ 0 & \text{otherwise} \end{cases}.$$

Algorithm 3 Diagonal-Thresholding (Johnstone and Lu, 2009)

Require: Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, sparsity level s

Ensure: vector $\hat{\theta} \in \mathbb{R}^p$

1: Let $\hat{\Sigma} = \frac{1}{n}(\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{X} - \bar{\mathbf{X}})$

2: Let $S \subseteq [p]$ contain the s coordinates of largest absolute value in $\text{diag}(\hat{\Sigma})$, Let $\hat{\Sigma}_S$ be the sub-matrix of $\hat{\Sigma}$ indexed by column and row support S.

3: Compute the top eigenvector $\hat{\mu}$ of $\hat{\Sigma}_S$, and the resulting DT estimator is the unit p-dimensional vector $\hat{\theta}$ with element

$$\hat{\theta}_j = \begin{cases} \text{the corresponding element in } \hat{\mu} & \text{if} \quad j \in S \\ 0 & \text{otherwise} \end{cases}.$$

Algorithm 4 Covariance-Thresholding (Krauthgamer et al., 2015)

Require: Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, sparsity level s

Ensure: vector $\hat{\theta} \in \mathbb{R}^p$

1: Let $\hat{\Sigma} = \frac{1}{n}(\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{X} - \bar{\mathbf{X}})$

2: Compute $T \in \mathbb{R}^{p \times p}$ by thresholding the entries of $\hat{\Sigma}$, namely,

$$T_{ij} = \begin{cases} \hat{\Sigma}_{ij} & \text{if} \quad |\hat{\Sigma}_{ij}| > t \\ 0 & \text{otherwise} \end{cases}.$$

3: Let $\hat{\mu} \in \mathbb{R}^p$ be the leading eigenvector of T

4: Let $S \subset [p]$ be the set of coordinates corresponding to the s largest absolute values in $\hat{\theta}$ and the resulting CT estimator is the unit p-dimensional vector $\hat{\theta}$

$$\hat{\theta}_j = \begin{cases} \frac{\hat{\mu}_j}{\|\hat{\mu}_S\|} & \text{if} \quad j \in S; \\ 0 & \text{otherwise} \end{cases}.$$

While these 3 algorithms focus on the estimation of ℓ_0-sparse PCA, they serve as good estimators of $\Delta\mu$ in (3.13). Naturally, one could solve the problem of sparse clustering by applying standard clustering algorithms (for example, standard K-Means and Gaussian Mixture Model) on the sub-data matrix with selected columns

indexed by the support of $\Delta\mu$. The performance of these three methods will be compared with that of our algorithms to be introduced in Section 3.3.

3.3 An iterative approach for sparse clustering

In this section, we will first introduce *Iterative 2-means*, which is designed to simultaneously do feature selection and sparse clustering for data generated from Model (3.11). We will also adapt Algorithms 2-4 to recover the support of $\Delta\mu$ for Model (3.13), and their performance will be compared with *Iterative 2-means*.

Algorithm 5 Iterative 2-means

Require: Design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, sparsity level s, number of iterations t

Ensure: subset $S \subseteq [p]$ of cardinality s

1: Let $m = \frac{p-s}{t}$, set $S = [p]$
2: **for** i in $\{1, ..., t\}$ **do**
3: Apply Lloyd's 2-means algorithm on X_S, obtain p-dimensional center vectors μ_0 and μ_1;
4: Update S to be the set of coordinates corresponding to the $[p - im]$ largest absolute values in $\mu_1 - \mu_0$
5: **end for**

We provide some intuition as to why we expect this algorithm to work. Recall that the objective of sparse clustering, as declared in (2.8), is to minimize the average within-cluster S-dissimilarity over $S \subseteq [p]$ and $C \in \mathcal{C}_\kappa^n$. That is to say, the goal is to simultaneously recover the important feature subset S and perform clustering on the sub-data matrix. The twin task is difficult, however, the sub-problems seem easier: if we know S in advance, this problem reduces to classical clustering which can be solved by, for example, K-means and GMM; and if we know the clustering results $C_1, \cdots, C_k$, this problem reduces to feature selection. Our *Iterative 2-means* is designed to tackle the latter problem, through an iterative approach.

Iterative 2-means adaptively recovers the important feature set step by step: in each step, we perform K means on current sub-data matrix and obtain two

p-dimensional center vectors μ_1 and μ_2; then a number of least important features are screened out by selecting the coordinates with the smallest separation between μ_1 and μ_2, and a sub-data matrix is returned for the next iteration. Notice that this algorithm has a parameter t, which is designed to be specified by users depending on the size of their dataset and available computing power. We suggest that for small p, one could simply let $t = p$; for large values of p, one could choose $\lfloor p/2 \rfloor$, or $\lfloor p/3 \rfloor, \cdots$, as t.

To demonstrate the utility and advantage of *Iterative 2-means*, we compare it with Algorithms 2-4 under the following set-up with a focus on sparse recovery. We generated n i.i.d. samples X_i from Model (3.11) with μ_0 the p-dimensional zero vector and μ_1 the p-dimensional vector of the form $\mu_1 = \sqrt{\lambda}\left(\frac{1}{\sqrt{s}}, \frac{1}{\sqrt{s}}, \ldots, \frac{1}{\sqrt{s}}, 0, 0, \ldots, 0\right)$. We assume the sparsity level s was a-priori known, and say that an execution of an algorithm is successful if it returns the support of μ exactly, i.e., if the output is the set $\{1, \ldots, s\}$. The *success rate* of an algorithm in M independent trials is the number of successful trials divided by M. In each experiment we fixed $n = p = 500$, and $\lambda = 4$ (signal strength here). We compare the performance of our *Iterative 2-means* to Algorithm 2-4 (SDP, DT, CT). For tuning parameters, we choose $m = 10$ in *Iterative 2-means* and $t = 5/\sqrt{n}$ in CT as suggested by Krauthgamer, Nadler, and Vilenchik (2015). As can be seen from Figure 3.1, *Iterative 2-means* clearly outperforms SDP, DT and CT. In Figure 3.2, we also plot the dot-products (in absolute value) between $\frac{\mu_1 - \mu_0}{\|\mu_1 - \mu_0\|}$ from *Iterative 2-Means* , and $\Delta\mu/\sqrt{\lambda}$ (normalized true $\Delta\mu$) under the same set-up as in Figure 3.1. As expected, the dot-product gets smaller as the sparsity s increases. For comparison, the figure also plots the dot products for CT, and clearly *Iterative 2-means* outperforms CT.

3.4 Discussion

In this Chapter, we first reviewed the literature on sparse PCA and establish the connection between sparse PCA to sparse clustering. Then we applied the aggregation estimator from Cai et al. (2013) to sparse clustering and provided its theoretical guarantee for the estimation of $\Delta\mu$ in (3.13) under the model (3.11).

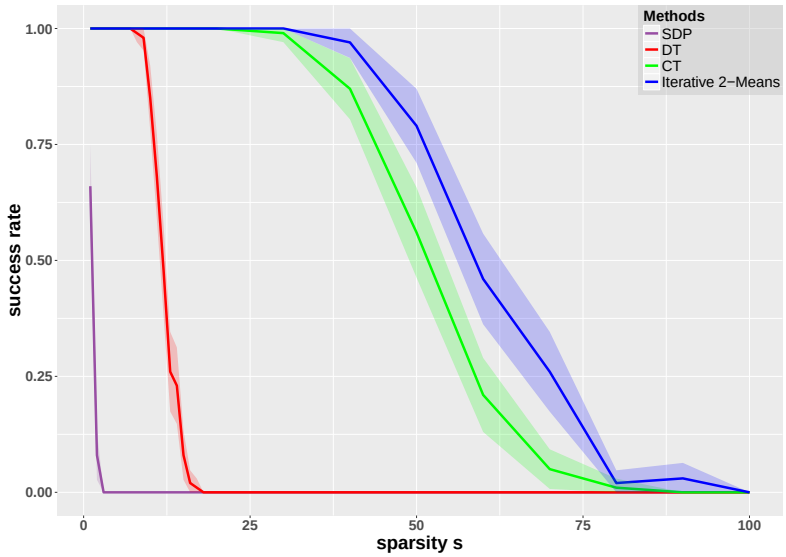


Figure 3.1: Success rate (with 95% confidence intervals) as a function of the sparsity for Algorithms 1-4 averaged over $M = 100$ runs, with $n = p = 500$.

We also adapted 3 computationally-efficient estimators developed in the context of sparse PCA to sparse clustering and proposed a new algorithm *Iterative 2-Means*. We compared these methods in sparse clustering with 2 clusters, and it shows that *Iterative 2-Means* outperforms the other 3 methods in terms of both sparse recovery and the estimation of $\Delta\mu$.

Notice that the study in this chapter only considered sparse clustering with 2 clusters and the within-cluster covariance is assumed to be Identity. What if we have k clusters with arbitrary different within-cluster covariance matrices? One could consider a Gaussian mixture model given by

$$\mathbf{X}_i \overset{\text{iid}}{=} \sum_{j=1}^{k} \alpha_j Z_j \in \mathbb{R}^p, \quad 1 \le i \le n, \tag{3.60}$$

where $Z_j \overset{\text{iid}}{\sim} N(\mu_j, \Sigma_j)$ for $1 \le j \le p$, $\alpha_1, \cdots, \alpha_k$ are dependent Bernoulli random variables ($\sum_{j=1}^{k} \alpha_i = 1$) with parameters $p_1, \cdots, p_k$, respectively. If we denote

$$\mathbf{\Delta} = [\Delta_1, \Delta_2, \cdots, \Delta_{k-1}] = [\mu_1 - \mu_k, \mu_2 - \mu_k, \cdots, \mu_{k-1} - \mu_k] \in \mathbb{R}^{p \times (k-1)}$$

and let $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_{k-1}]^T$ and $C = \text{cov}(\alpha)$, then

$$\text{Cov}(\mathbf{X}_1) = \mathbf{\Delta} C \mathbf{\Delta}' + \sum_{j=1}^{k} p_j \Sigma_j.$$

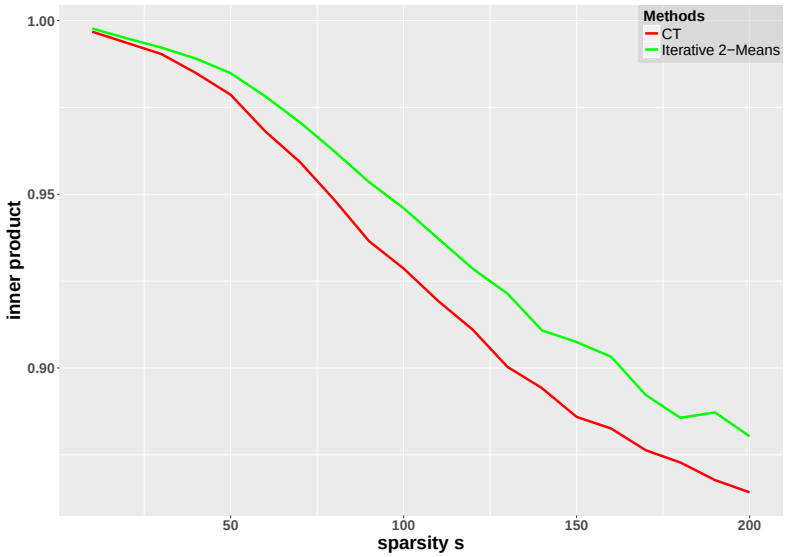


Figure 3.2: Comparison of *Iterative 2-Means* and CT (Algorithm 4) for $n = p = 500$, $\lambda = 4$, averaged over 100 runs. The dots are the average of the absolute dot products between $\hat{\mu}$, the output of *Iterative 2-Means* and CT, and the true μ.

Since C is not diagonal, even if we assume that $\Sigma_j = I_p$ for all j, estimating $\mathbf{\Delta}$ via estimating the principal subspace spanned by the $k-1$ leading eigenvectors of $\mathrm{Cov}(\mathbf{X}_1)$ could still be problematic. Thus, when $k \geq 3$, the problem is not as directly related to sparse PCA as it was shown to be when $k = 2$. In Chapter 4, we will propose a simple approach to sparse clustering, which can deal with arbitrary number of clusters.

Chapter 4

Sparse Alternate Sum Clustering

In this chapter, we consider the problem of sparse clustering, where it is assumed that only a subset of the features are useful for clustering purposes. In the framework of the COSA method of Friedman and Meulman (2004), subsequently improved in the form of the Sparse K-means method of Witten and Tibshirani (2010), a natural and simpler hill-climbing approach is introduced. The new method is shown to be competitive with these two methods and others.

4.1 The Algorithm

Hill-climbing methods are iterative in nature, making 'local', that is, 'small' changes at each iteration. They have been studied in the context of graph partitioning, e.g., by Kernighan and Lin (1970) and Carson and Impagliazzo (2001), among others. In the context of sparse clustering, we find the K-medoids variant of Aggarwal et al. (1999), which includes a hill-climbing step. Many of the methods cited in Section 2.1 use alternate optimization in some form (e.g., EM), which can be interpreted as hill-climbing. Our method to be presented in this chapter, is instead directly formulated as a hill-climbing approach, making it simpler and, arguably, more principled than COSA or Sparse K-means.

4.1.1 Our approach: SAS Clustering

Let $\hat{C}$ be an algorithm for clustering based on dissimilarities. Formally, $\hat{C} : \mathbb{D} \times \mathbb{N} \mapsto \mathcal{C}$, where $\mathbb{D}$ is a class of dissimilarity matrices and $\mathcal{C} := \bigcup_n \bigcup_\kappa \mathcal{C}_\kappa^n$, and for $(\delta, \kappa) \in \mathbb{D} \times \mathbb{N}$ with δ of dimension n, $\hat{C}(\delta, \kappa) \in \mathcal{C}_\kappa^n$. Note that $\hat{C}$ could be a hill-climbing method for graph partitioning, or K-medoids (or K-means if we are provided with points in a vector space rather than dissimilarities), or a spectral method, namely, any clustering algorithm that applies to dissimilarities. (In this chapter, we will use K-means for numerical data and K-medoids for categorical data using hamming distances as dissimilarities.) For $S \subset [p]$, define

$$\boldsymbol{\delta}_S = (\delta_a(i,j) : a \in S; i,j \in [n]) \quad \text{and} \quad \boldsymbol{\delta} = \boldsymbol{\delta}_{[p]}. \tag{4.1}$$

Our procedure is described in Algorithm 6.

Algorithm 6 Sparse Alternate Similarity (SAS) Clustering

Input: dissimilarities $(\delta_a(i,j) : a \in [p], i,j \in [n])$, number of clusters κ, number of features s

Output: feature set S, group assignment function C

Initialize: For each $a \in [p]$, compute $C_a \leftarrow \hat{C}(\boldsymbol{\delta}_a, \kappa)$ and then $\Delta_a[C_a]$. Let $S \subset [p]$ index the smallest s among these.

Alternate between the following steps until 'convergence':

1: Keeping S fixed, compute $C \leftarrow \hat{C}(\boldsymbol{\delta}_S, \kappa)$.

2: Keeping C fixed, compute $S \leftarrow \arg\min_{|S|=s} \Delta_S[C]$.

The use of algorithm $\hat{C}$ in Step 1 is an attempt to minimize $C \mapsto \Delta_S[C]$ over $C \in \mathcal{C}_\kappa^n$. The minimization in Step 2 is over $S \subset [p]$ of size s and it is trivial. Indeed, the minimizing S is simply made of the s indices $a \in [p]$ corresponding to the smallest $\Delta_a[C]$. For the choice of parameters κ and s, any standard method for tuning parameters of a clustering algorithm applies, for example, by optimization of the gap statistic of Tibshirani et al. (2001). We note that the initialization phase, by itself, is a pure coordinate-wise approach that has analogs in the Euclidean setting as mentioned Section 2.1.3. The hill-climbing process is the iteration phase.

Remark 2. We tried another initialization in Algorithm 6 consisting of drawing a feature set S at random. We found that the algorithm behaved similarly. (Results not reported here.)

Compared with COSA and Sparse K-means, and other methods based on penalties, we note that the choice of features in our SAS algorithm is much simpler, using a hill-climbing approach instead.

4.1.2 Number of iterations needed

A first question of interest is whether the iterations improve the purely coordinate-wise method, defined as the method that results from stopping after one pass through Steps 1-2 in Algorithm 6 (no iteration). Although this is bound to vary with each situation, we examine an instance where the data come from the mixture of three Gaussians with sparse means. In detail, the setting comprises 3 clusters with 30 observations each and respective distributions $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$, $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathcal{N}(-\boldsymbol{\mu}, \mathbf{I})$, with $\boldsymbol{\mu} = (\mu, \ldots, \mu, 0, \ldots, 0)$ having 50 μ's and 450 zeros. We assume that $\kappa = 3$ and $s = 50$ are both given, and we run the SAS algorithm and record the Rand indexes (Rand, 1971) and symmetric differences $|S_* \triangle \hat{S}|$ as the end of each iteration of Steps 1-2. The setting is repeated 400 times. The means and confidence intervals under different regimes ($\mu = 0.6$, $\mu = 0.7$, $\mu = 0.8$, $\mu = 0.9$) are shown in Figure 4.1. At least in this setting, the algorithm converges in a few iterations and, importantly, these few iterations bring significant improvements, particularly over the purely coordinate-wise algorithm.

4.1.3 Selection of the sparsity parameter

We consider the problem of selecting κ, the number of clusters, as outside of the scope of this work, as it is intrinsic to the problem of clustering and has been discussed extensively in the literature — see (Kou, 2014; Tibshirani et al., 2001) and references therein. Thus we assume that κ is given. Besides κ, our algorithm has one tuning parameter, the sparsity parameter s, which is the number of useful features for clustering, meaning, the cardinality of set S in (2.8).

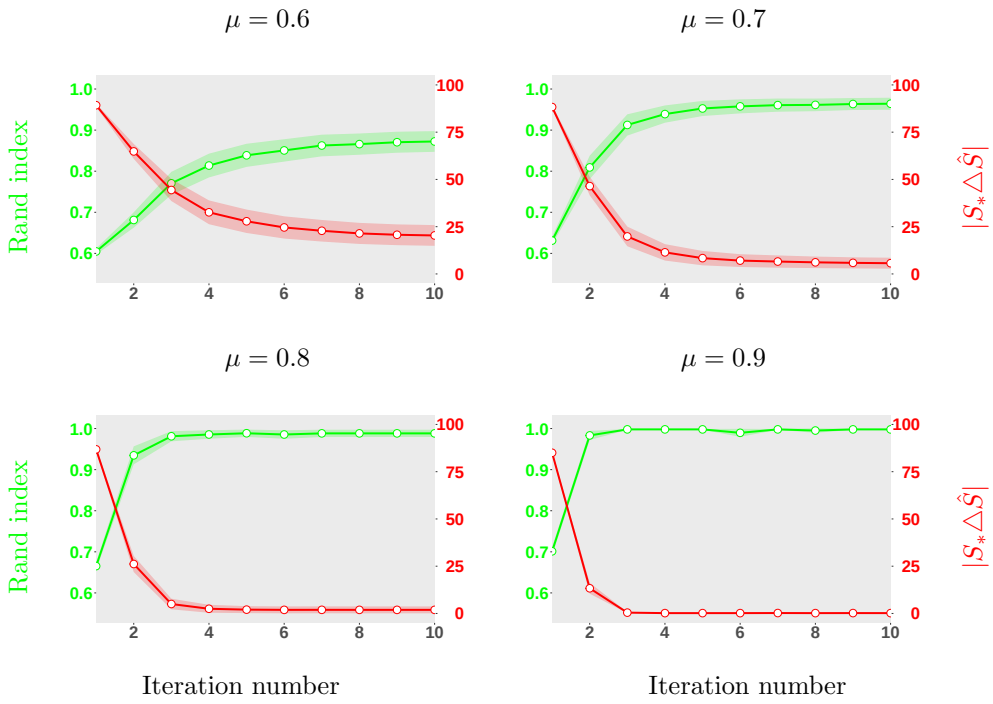


Figure 4.1: Means (and 95% confidence intervals of the means) of Rand indexes and symmetric differences.

Inspired by the gap statistic of Tibshirani et al. (2001), which was designed for selecting the number of clusters κ in standard K-means clustering, we propose a permutation approach for selecting s. Let Δ_s^{obs} denote the average within-cluster dissimilarity of the clustering computed by the algorithm on the original data with input number of features s. Let Δ_s^{perm} denote the same quantity but obtained from a random permutation of the data — a new sample is generated by independently permuting the observations within each feature. The gap statistic (for s) is then defined as

$$\text{gap}(s) = \log \Delta_s^{\text{obs}} - \mathbb{E}(\log \Delta_s^{\text{perm}}). \tag{4.2}$$

In practice, the expectation is estimated by Monte Carlo, generating B random permuted datasets. A large gap statistic indicates a large discrepancy between the observed amount of clustering and that expected of a null model (here a permutation of the data) with no salient clusters.

The optimization of the gap statistics over $s \in [p]$ is a discrete optimization

problem. An exhaustive search for s would involve computing p gap statistics, each requiring B runs of the SAS algorithm. This is feasible when p and B are not too large.[1] See Algorithm 7, which allows for coarsening the grid.

Algorithm 7 SAS Clustering with Grid Search

Input: Dissimilarities $(\delta_a(i,j) : a \in [p], i,j \in [n])$, number of clusters κ, step size h, number of Monte Carlo permutations B

Output: Number of useful features $\hat{s}$, feature set S, group assignment C

for $s = 1$ to p with step size h **do**

Run **Algorithm 6** to get the feature set S_s and group assignment C_s

Run **Algorithm 6** on B permuted datasets to get the gap statistic G_s

end for

return Let $\hat{s} = \arg\max_s G_s$ and return $S_{\hat{s}}$ and $C_{\hat{s}}$

To illustrate the effectiveness of choosing s using the gap statistic, we computed the gap statistic for all $s \in [p]$ in the same setting as that of Section 4.1.2 with $\mu = 1$. The result of the experiment is reported in Figure 4.2. Note that, in this relatively high SNR setting, the gap statistic achieves its maximum at the correct number of features.

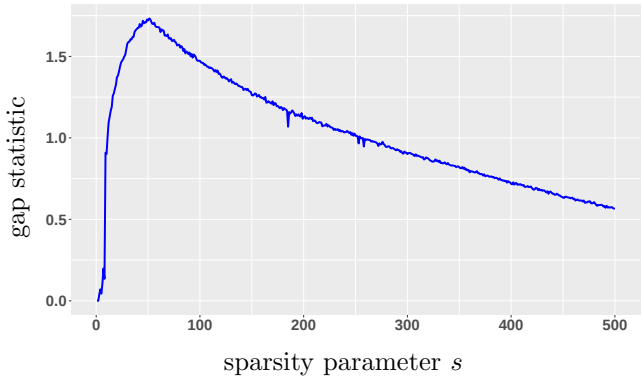


Figure 4.2: A plot of the gap statistic for each $s \in [p]$ for a Gaussian mixture with 3 components (30 observations in each cluster) in dimension $p = 500$.

In this experiment, at least, the gap statistic seems unimodal (as a function of

[1] In our experiments, we choose $B = 25$ as in the code that comes with (Witten and Tibshirani, 2010).

s). If it were the case, we could use a golden section search, which would be much faster than an exhaustive grid search.

4.2 Numerical experiments

We performed a number of numerical experiments, both on simulated data and on real (microarray) data to compare our method with other proposals. Throughout this section, we standardize the data coordinate-wisely, we assume that the number of clusters is given, and we use the gap statistic of Tibshirani et al. (2001) to choose the tuning parameter s in our algorithm.

4.2.1 A comparison of SAS Clustering with Sparse K-means and IF-PCA-HCT

We compare our Algorithm 6 with IF-PCA-HCT (Jin and Wang, 2014) and Sparse K-means (Witten and Tibshirani, 2010) in the setting of Section 4.1.2. We note that IF-PCA-HCT was specifically designed for that model and that Sparse K-means was shown to numerically outperform a number of other approaches, including standard K-means, COSA (Friedman and Meulman, 2004), model-based clustering (Raftery and Dean, 2006), the penalized log-likelihood approach of (Pan and Shen, 2007) and the classical PCA approach. We use the gap statistic to tune the parameters of SAS Clustering and Sparse K-means. (SAS_gs uses a grid search while SAS_gss uses a golden section search.) IF-PCA-HCT is tuning-free — it employs the higher criticism to automatically choose the number of features.

In Table 4.1a, we report the performance for these three methods in terms of Rand index (Rand, 1971) for various combinations of μ and p. Each situation was replicated 50 times. As can be seen from the table, SAS Clustering outperforms IF-PCA-HCT, and performs at least as well as Sparse K-means and sometimes much better (for example when $p = 500$ and $\mu = 0.7$). We examine a dataset from this situation in depth, and plot the weights resulted from Sparse K-means on this dataset, see Figure 4.3. As seen in this figure, and also as mentioned in (Witten and Tibshirani, 2010), Sparse K-means generally results in more features

with non-zero weights than the truth. These extraneous features, even with small weights, may negatively impact the clustering result. In this specific example, the Rand index from Sparse K-means is 0.763 while our approach gives a Rand index of 0.956. Let $S_* \subset [p]$ denote the true feature set and $\hat{S}$ the feature set that our method return. In this example, $|S_* \triangle \hat{S}| = 12$.

While both SAS Clustering and Sparse K-means use the gap statistic to tune the parameters, IF-PCA-HCT tunes itself analytically without resorting to permutation or resampling, and (not surprisingly) has the smallest computational time among these three methods. However, as can be seen from Table 4.1a, the clustering results given by IF-PCA-HCT are far worse than those resulted from the other two methods. In Table 4.1b, we report the performance of SAS Clustering and Sparse K-means in terms of the running time, under the same setting as that in Table 4.1a but with tuning parameters for both of the methods given (so that the comparisons are fair). As can be seen in Table 4.1b, SAS Clustering shows a clear advantage over Sparse K-means in terms of the running time, and as p increases, the advantage becomes more obvious. (Note that both SAS and Sparse K-means are implemented in R code and, in particular, the code is not optimized.)

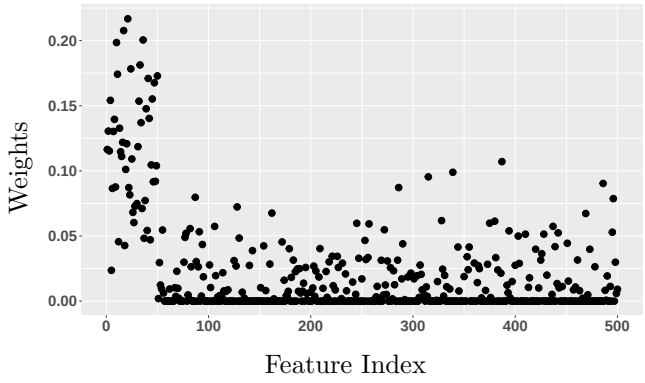


Figure 4.3: A typical example of the weights that Sparse K-means returns.

4.2.2 A more difficult situation (same covariance)

In Section 4.2.1, the three groups had identity covariance matrix. In this section, we continue comparing our approach with Sparse K-means and IF-PCA-HCT

Table 4.1a: Comparison results for the simulations in Section 4.2.1. The reported values are the mean (and sample standard deviation) of the Rand indexes over 50 simulations.

μ	methods	p = 100	p = 200	p = 500	p = 1000
0.6	SAS_gs	0.907 (0.048)	0.875 (0.066)	0.827 (0.076)	0.674 (0.096)
	SAS_gss	0.900 (0.054)	0.860 (0.066)	0.781 (0.008)	0.701(0.050)
	Sparse K	0.886 (0.068)	0.807 (0.064)	0.744 (0.046)	0.704 (0.043)
	IF-PCA	0.664(0.042)	0.645(0.051)	0.605 (0.045)	0.593(0.038)
0.7	SAS_gs	0.953 (0.030)	0.965 (0.028)	0.960 (0.032)	0.855 (0.102)
	SAS_gss	0.953 (0.031)	0.961 (0.031)	0.921 (0.088)	0.789 (0.104)
	Sparse K	0.942 (0.045)	0.915 (0.071)	0.802 (0.087)	0.790 (0.087)
	IF-PCA	0.681(0.036)	0.653(0.044)	0.629(0.057)	0.614(0.055)
0.8	SAS_gs	0.986 (0.020)	0.985 (0.022)	0.987 (0.016)	0.966 (0.052)
	SAS_gss	0.984 (0.020)	0.983 (0.019)	0.987 (0.0178)	0.892 (0.122)
	Sparse K	0.985 (0.020)	0.975 (0.029)	0.961 (0.07)	0.948 (0.074)
	IF-PCA	0.691(0.043)	0.675(0.056)	0.639(0.068)	0.623(0.059)
0.9	SAS_gs	0.997 (0.008)	0.997 (0.008)	0.997 (0.007)	0.995 (0.010)
	SAS_gss	0.996 (0.010)	0.996 (0.009)	0.997 (0.009)	0.969 (0.076)
	Sparse K	0.996 (0.010)	0.992 (0.013)	0.992(0.016)	0.993 (0.013)
	IF-PCA	0.700(0.031)	0.682(0.051)	0.654(0.057)	0.627(0.065)
1.0	SAS_gs	0.999 (0.005)	1.000 (0.003)	1.000 (0.003)	0.999 (0.004)
	SAS_gss	0.998 (0.007)	1.000 (0.003)	1.000 (0.004)	0.998 (0.006)
	Sparse K	0.998 (0.007)	0.999 (0.005)	0.996 (0.010)	0.996 (0.009)
	IF-PCA	0.717(0.034)	0.710(0.039)	0.659(0.063)	0.639(0.060)

under a more difficult situation, where each of the 3 clusters have 30 points sampled from different p-variate normal distributions ($p = 100, 200, 500, 1000$), with different mean vectors

$$\boldsymbol{\mu}_1 = [1.02, 1.04, ..., 2, \underbrace{0, ..., 0}_{p-50 \text{ zeros}}],$$

$$\boldsymbol{\mu}_2 = [1.02 + \delta_\mu, 1.04 + \delta_\mu, ..., 2 + \delta_\mu, \underbrace{0, ..., 0}_{p-50 \text{ zeros}}],$$

$$\boldsymbol{\mu}_3 = [1.02 + 2\delta_\mu, 1.04 + 2\delta_\mu, ..., 2 + 2\delta_\mu, \underbrace{0, ..., 0}_{p-50 \text{ zeros}}],$$

and same diagonal covariance matrix $\boldsymbol{\Sigma}$ across groups, a random matrix with eigenvalues in $[1, 5]$. We used 50 repeats and varied δ_μ from 0.6 to 1.0. The results

Table 4.1b: Comparison of running time of SAS Clustering (with the number of features s given) and Sparse K-means (with known tuning parameter s in (2.13)) in the setting of Section 4.2.1. Reported is the averaged running time (in seconds) over 100 repeats, with sample standard deviation in parentheses.

δ_μ	methods	p = 100	p = 200	p = 500	p = 1000
0.6	SAS	0.086 (0.031)	0.130 (0.044)	0.217 (0.088)	0.271 (0.118)
	Sparse K	0.113 (0.034)	0.220 (0.053)	0.445 (0.101)	0.850 (0.156)
0.7	SAS	0.077 (0.021)	0.104 (0.027)	0.207 (0.085)	0.316 (0.147)
	Sparse K	0.107 (0.028)	0.235 (0.057)	0.471 (0.123)	0.945 (0.194)
0.8	SAS	0.056 (0.019)	0.088 (0.022)	0.182 (0.062)	0.313 (0.118)
	Sparse K	0.091 (0.029)	0.213 (0.051)	0.574 (0.134)	0.984 (0.262)
0.9	SAS	0.055 (0.017)	0.080 (0.024)	0.136 (0.048)	0.289 (0.131)
	Sparse K	0.094 (0.023)	0.196 (0.052)	0.482 (0.101)	0.982 (0.261)
1.0	SAS	0.051(0.012)	0.089 (0.021)	0.146 (0.045)	0.272 (0.095)
	Sparse K	0.095(0.019)	0.186 (0.044)	0.554 (0.107)	1.225 (0.270)

are reported in Table 4.2a. We see there that, in this setting, our method is clearly superior to Sparse K-means and IF-PCA-HCT. We also report the symmetric difference $|S_* \triangle \hat{S}|$ between the estimated feature set $\hat{S}$ and the true feature set S_*, as can be seen in Table 4.2b. Our algorithm is clearly more accurate in terms of feature selection.

4.2.3 A more difficult situation (different covariances)

In both Section 4.2.1 and Section 4.2.2, the three groups have the same covariance matrix. In this section, we continue comparing our approach with Sparse K-means and IF-PCA-HCT under an even more difficult situation, where the mean vectors are the same as in Section 4.2.2 with $\delta_\mu = 1.0$, but now the covariances are different: $\mathbf{\Sigma}_1$, $\mathbf{\Sigma}_2$ and $\mathbf{\Sigma}_3$ are random matrices with eigenvalues in $[1, 2]$, $[2, 3]$ and $[3, 4]$, respectively. We used 50 repeats in this simulation. The results, reported in Table 4.3, are consistent with the results of Section 4.2.2: our method clearly outperforms Sparse K-means and IF-PCA-HCT, both in terms of clustering and feature selection.

Notice that the 3 clusters are well separated in the first 50 features as can be

Table 4.2a: Comparison of SAS Clustering with Sparse K-means and IF-PCA in the setting of Section 4.2.2 in terms of Rand index. Reported is the averaged Rand index over 50 repeats, with the standard deviation in parentheses.

δ_μ	methods	p = 100	p = 200	p = 500	p = 1000
0.6	SAS_gs	0.718 (0.037)	0.702 (0.037)	0.611 (0.044)	0.574 (0.027)
	SAS_gss	0.714 (0.028)	0.692 (0.038)	0.635 (0.044)	0.595(0.026)
	Sparse K	0.590 (0.030)	0.594 (0.034)	0.595 (0.034)	0.571 (0.023)
	IF-PCA	0.619(0.037)	0.590(0.037)	0.572 (0.024)	0.564(0.020)
0.8	SAS_gs	0.852 (0.047)	0.844 (0.052)	0.797 (0.066)	0.670 (0.082)
	SAS_gss	0.848 (0.050)	0.819 (0.070)	0.752 (0.060)	0.686 (0.043)
	Sparse K	0.662 (0.057)	0.646 (0.063)	0.657 (0.062)	0.639 (0.054)
	IF-PCA	0.646(0.040)	0.634(0.047)	0.603(0.046)	0.575(0.040)
1.0	SAS_gs	0.940 (0.035)	0.947 (0.033)	0.941 (0.037)	0.919 (0.065)
	SAS_gss	0.935 (0.037)	0.941 (0.038)	0.922 (0.059)	0.799 (0.099)
	Sparse K	0.798 (0.085)	0.814 (0.078)	0.742 (0.080)	0.708 (0.070)
	IF-PCA	0.677(0.041)	0.644(0.056)	0.618(0.052)	0.604(0.047)

seen from the construction of the data, but when 450 noise features are present in the datasets, the task of clustering becomes difficult. See Figure 4.4(b) as an example where we project a representative dataset onto the first two principal components of the whole data matrix. However, if we are able to successfully select out the first 50 features and apply classical clustering algorithms, then we are able to achieve better results. See Figure 4.4(a), where we project the same dataset onto the first two principal components of the data submatrix consisting of the first 50 columns (features). To illustrate the comparisons, we also plot in Figure 4.4 the clustering results by these three methods.

4.2.4 Clustering non-euclidean data

In the previous simulations, all the datasets were Euclidean. In this section, we apply our algorithm on categorical data (with Hamming distance) and compare its performance with Sparse K-medoids[2]. In this example, we generate 3 clusters with 30 data points each from three different distributions on the Hamming space

[2]We modified the function of Sparse K-means in the R package 'sparcl', essentially replacing K-means with K-medoids, so that it can be used to cluster categorical data.

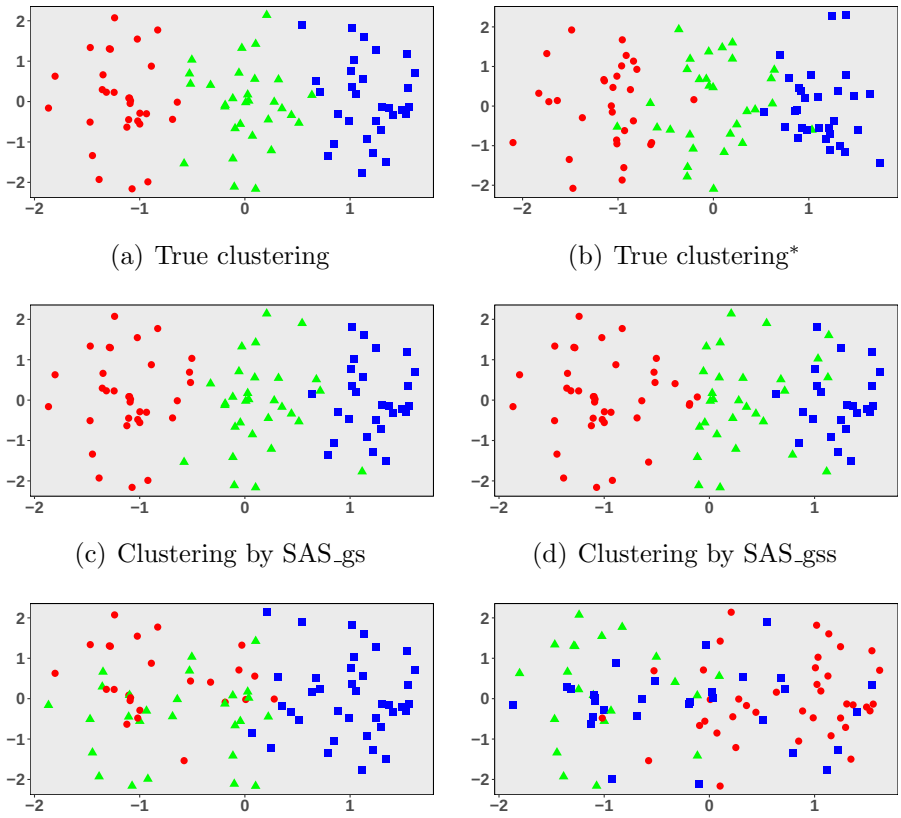
Table 4.2b: Comparison of SAS Clustering with Sparse K-means and IF-PCA in the setting of Section 4.2.2 in terms of feature selection. Reported is the averaged symmetric difference over 50 repeats, with the standard deviation in parentheses.

δ_μ	methods	p = 100	p = 200	p = 500	p = 1000
0.6	SAS_gs	26.3(4.7)	37.7 (7.9)	86.2 (20.5)	121.0(28.3)
	SAS_gss	27.9(5.3)	44.1 (12.1)	100.5 (43.3)	143.1(57.0)
	Sparse K	43.8 (7.3)	86.8(35.3)	163.9(105.9)	170.6 (124.6)
	IF-PCA	49.4(3.8)	72.3(15.8)	129.7 (61.4)	185.8(126.3)
0.8	SAS_gs	17.4(3.9)	19.0(3.9)	31.7 (17.2)	94.2 (48.3)
	SAS_gss	17.7 (4.6)	21.9 (6.5)	57.9 (43.7)	132.9 (85.0)
	Sparse K	28.5(13.2)	63.4(28.5)	163.6 (102.9)	218.6 (129.9)
	IF-PCA	50.8(5.2)	75.4(16.5)	126.9(61.2)	209.3(130.5)
1.0	SAS_gs	10.5 (3.7)	10.2 (3.4)	12.4 (3.7)	22.7(19.6)
	SAS_gss	11.7(3.9)	12.5 (4.1)	17.1 (12.8)	100.2 (84.8)
	Sparse K	13.6 (10.8)	49.7 (38.2)	204.88 (104.5)	265 (165.1)
	IF-PCA	49.3(4.0)	67.9(14.1)	124.8(53.3)	226.3(146.7)

Table 4.3: Comparison of SAS Clustering with Sparse K-means and IF-PCA in the setting of Section 4.2.3. Reported are the Rand index and symmetric difference, averaged over 50 repeats. The standard deviations are in parentheses.

Method	SAS_gs	SAS_gss	Sparse K-means	IF-PCA
Rand index	0.920 (0.054)	0.858 (0.098)	0.710 (0.022)	0.668 (0.041)
$\lvert S_* \triangle \hat{S} \rvert$	8.7 (3.8)	13.0 (7.9)	297.2 (75.6)	118.6 (56.8)

of dimension p. Each distribution is the tensor product of Bernoulli distributions with success probabilities $q_a \in [0,1]$ for $a \in [p]$. For the first distribution, $q_a = q$ for $1 \leq a \leq 5$ and $q_a = 0.1$ otherwise. For the second distribution, $q_a = q$ for $6 \leq a \leq 10$ and $q_a = 0.1$ otherwise. For the third distribution, $q_a = q$ for $11 \leq a \leq 15$ and $q_a = 0.1$ otherwise. See Table 4.4, where we compare these two methods in terms of Rand index for various combination of q and p. Each situation was replicated 50 times. As can be seen from the table, SAS Clustering significantly outperforms Sparse K-medoids in most situations. We examined why, and it turns out that Sparse K-medoids works well if the tuning parameter s in equation (2.13) is given, but it happens that the gap statistic often fails to give a good estimate of s in this categorical setting. We are not sure why.



(a) True clustering (b) True clustering*

(c) Clustering by SAS_gs (d) Clustering by SAS_gss

(e) Clustering by Sparse K-means (f) Clustering by IF-PCA

Figure 4.4: Projection of a dataset from Section 4.2.3 onto the first two principal components of the data submatrix, where only the first 50 columns are kept. *Different from the other 5 subfigures, here the data points are projected onto the first two principal components of the whole data matrix.

4.2.5 Comparisons as the number of clusters κ increases

In Sections 4.2.1 – 4.2.4, we have fixed the number of clusters to be 3 and considered the effects of cluster separation (μ, q), sparsity (p) and cluster shape (Identity covariance, same and different covariance matrices across groups) in the comparisons. In this section, we continue to compare our approach with Sparse K-means and IF-PCA-HCT as the number of clusters κ increases from 2 to 10. The set-up here is different from the above sections. We sample κ sub-centers from a 50-variate normal distribution $\mathcal{N}(\mathbf{0}, 0.4 \times \mathbf{I}_{50})$[3] and concatenate each of the sub-centers with 450 zeros to have κ random centers, $\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_\kappa$, of length

[3]The constant 0.4 was chosen to make the task of clustering neither too easy nor too difficult.

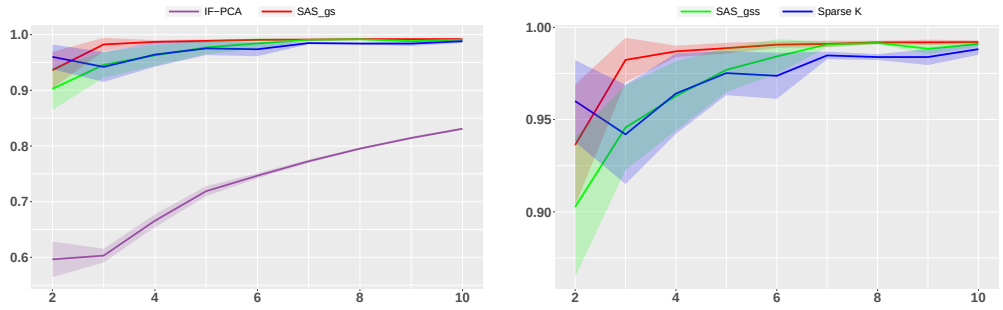
Table 4.4: Comparison results for Section 4.2.4. The reported values are the mean (and standard error) of the Rand indexes over 50 simulations.

q	methods	p = 30	p = 60	p = 100	p = 200
0.6	SAS_gs	0.878 (0.060)	0.872 (0.042)	0.864 (0.057)	0.863 (0.053)
	Sparse K-medoids	0.694 (0.045)	0.663 (0.054)	0.654 (0.049)	0.639 (0.044)
0.7	SAS_gs	0.954 (0.023)	0.960 (0.026)	0.942 (0.026)	0.948 (0.033)
	Sparse K-medoids	0.807 (0.126)	0.763 (0.077)	0.716 (0.060)	0.686 (0.062)
0.8	SAS_gs	0.989 (0.011)	0.984 (0.019)	0.983 (0.019)	0.978 (0.021)
	Sparse K-medoids	0.946 (0.090)	0.889 (0.099)	0.846 (0.100)	0.787 (0.093)
0.9	SAS_gs	0.998 (0.005)	0.999 (0.003)	0.997 (0.007)	0.997 (0.006)
	Sparse K-medoids	0.997 (0.006)	0.994 (0.036)	0.983 (0.044)	0.966 (0.065)

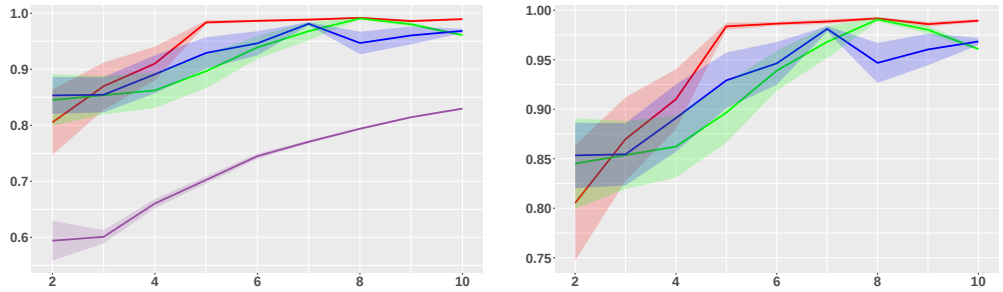
500, which carry at least 450 noise features. Once the centers are generated, we construct κ clusters with 30 (20 in the second set-up) observations each, sampled from respective distributions $\mathcal{N}(\boldsymbol{\mu}_i, \mathbf{I}_{500})$ with $i = 1, 2, \cdots, \kappa$. Each setting is repeated 50 times. The means and confidence intervals with different κ's are shown in Figure 4.5(a) and Figure 4.5(b). Once again, the results were consistent with earlier results in that SAS Clustering outperforms IF-PCA-HTC and performs at least as well as Sparse K-means with different κ's. We also notice that the clustering results given by all these three methods become better as κ increases. This can be explained by the increased effective sample sizes ($30 \times \kappa$ or $20 \times \kappa$) as κ increases.

4.2.6 Applications to gene microarray data

We compare our approach with others on real data from genetics. Specifically, we consider the same microarray datasets (listed in Table 4.5) used by Jin and Wang (2014) to evaluate their IF-PCA method. Each of these 10 data sets consists of measurements of expression levels of p genes in n patients from κ different classes (e.g., normal, diseased). We notice from Table 4.5 that p is much greater than n, illustrating a high-dimensional setting. We also mention that, although the true labels are given by the groups the individuals belong to, they are only used as the *ground truth* when we report the classification errors of the different methods in



(a) 30 observations in each cluster



(b) 20 observations in each cluster

Figure 4.5: Comparison of SAS Clustering with Sparse K-means and IF-PCA in the setting of Section 4.2.5. Reported are the means (and confidence intervals) of Rand indexes (y-axis) as the number of clusters, κ (x-axis), increases. For each sub-figure, we separately put the same plot on the right with the results of SAS Clustering and Sparse K-means only, which clearly outperform IF-PCA.

Table 4.6. For detailed descriptions and the access to these 10 datasets, we refer the reader to (Jin and Wang, 2014).

In Table 4.6, we report the classification errors of 10 different methods on these datasets. Among these 10 methods, the results from K-means, K-means++ (Arthur and Vassilvitskii, 2007), hierarchical clustering, SpectralGem (Lee et al., 2010) and IF-PCA-HCT (Jin and Wang, 2014) are taken from (Jin and Wang, 2014). We briefly mention that K-means++ is Lloyd's algorithm for K-means but with a more careful initialization than purely random; hierarchical clustering is applied to the normalized data matrix X directly without feature selection; and SpectralGem is PCA-type method. In addition to these 5 methods, we also include 3 other methods: AHP-GMM (Wang and Zhu, 2008), which is an adaptively hierarchically penalized Gaussian-mixture-model based clustering method, Regu-

larized K-means (Sun et al., 2012), and Sparse K-means (Witten and Tibshirani, 2010).

We can offer several comments. First, our method is overall comparable to Sparse K-means and IF-PCA, which in general outperform the other methods. It is interesting to note that SAS_gss outperforms SAS_gs on a couple of datasets. However, we caution the reader against drawing hard conclusions based on these numbers, as some of the datasets are quite small. For example, the Brain dataset has $\kappa = 5$ groups and a total sample size of $n = 42$, and is very high-dimensional with $p = 5{,}597$. Second, for Breast Cancer, Prostate Cancer, SRBCT and Su-Cancer, all methods perform poorly with the best error rate exceeding 31%. However, we note that even when the task is classification where class labels in the training sets are given, these data sets are still hard for some well-known classification algorithms (Dettling, 2004; Yousefi et al., 2010). Third, we notice that in (Sun et al., 2012), clustering results of the Leukemia and Lymphoma datasets have also been compared. The error rate on Lymphoma given by Regularized K-means in (Sun et al., 2012) is the same as reported here, however, the error rate on Leukemia is smaller than the result reported here. This is due to the fact that they applied preprocessing techniques to screen out some inappropriate features and also imputed the missing values using 5 nearest neighbors on this data set. Interestingly, Wang and Zhu (2008) also reported a better error rate on SRBCT data using their AHP-GMM method. However, they split the data into training set and testing set, fit the penalized Gaussian mixture model and report the training error and testing error respectively.

4.3 Discussion

In this chapter, we presented a simple method for feature selection in the context of sparse clustering. The method is arguably more natural and simpler to implement than COSA or Sparse K-means. At the same time, it performs comparably or better than these methods, both on simulated and on real data.

At the moment, our method does not come with any guarantees, other than

Table 4.5: 10 gene microarray datasets.

#	Data Name	κ	p	n (with sample size from each cluster)
1	Brain	5	5597	42 (10+10+10+4+8)
2	Breast	2	22215	276 (183+93)
3	Colon	2	2000	62 (22+40)
4	Lung	2	12533	181 (150+31)
5	Lung(2)	2	12600	203 (139+64)
6	Leukemia	2	3571	72 (47+25)
7	Lymphoma	3	4026	62 (42+9+11)
8	Prostate	2	6033	102 (50+52)
9	SRBCT	4	2308	63 (23+8+12+20)
10	SuCancer	2	7909	174 (83+91)

that of achieving a local minimum if the iteration is stopped when no improvement is possible. Just like other iterative methods based on alternating optimization, such as Lloyd's algorithm for K-means, proving a convergence to a good local optimum (perhaps even a global optimum) seems beyond reach at the moment. COSA and Sparse K-means present similar challenges and have not been analyzed theoretically. IF-PCA has some theoretical guarantees developed in the context of a Gaussian mixture model (Jin and Wang, 2014) — see also Jin et al. (2015). More theory for sparse clustering is developed in (Azizyan et al., 2013; Chan and Hall, 2010; Verzelen and Arias-Castro, 2014).

Acknowledgement

This chapter, in full, has been published in Computational Statistics and Data Analysis. Ery Arias-Castro, Xiao Pu, "A Simple Approach to Sparse Clustering", *Computational Statistics and Data Analysis*,105 (2017): 217-228. The dissertation author is the corresponding author of this material.

Table 4.6: Comparison of SAS Clustering with other clustering methods on 10 gene microarray datasets. (In **bold** is the best performance.)

Data set	K-means	K-means++	Hier	SpecGem	IF-PCA	AHP-GMM	RKmeans	Sparse K	SAS_{gs}	SAS_{gss}
Brain	.286	.472	.524	**.143**	.262	.214	.262	.190	.310	.310
Breast	.442	.430	.500	**.438**	.406	.460	.442	.449	.485	.445
Colon	.443	.460	.387	.484	.403	**.129**	.355	.306	**.129**	.403
Lung	.116	.196	.177	.122	**.033**	.116	.094	.122	.099	.099
Lung(2)	.436	.439	.301	.434	**.217**	.438	**.217**	.315	.315	.315
Leukemia	.278	.257	.278	.292	.069	**.028**	.347	**.028**	**.028**	**.028**
Lymphoma	.387	.317	. 468	.226	.065	.484	**.016**	**.016**	**.016**	**.016**
Prostate	.422	.432	.480	.422	.382	.422	.441	**.373**	.431	.431
SRBCT	.556	.524	.540	.508	.444	.476	.556	**.317**	.460	.365
SuCancer	.477	.459	.448	.489	**.333**	.477	.477	.477	.483	.483

Chapter 5

Semiparametric Estimation of Symmetric Mixture Models

In this chapter, we consider fitting the mixture model (2.23) in a more general setting, where the mixture components may have different shape:

$$g(x) = \sum_{j=1}^{k} \pi_j f_j(x - \mu_j), \quad \sum_{j=1}^{k} \pi_j = 1, \quad x \in \mathbb{R}. \tag{5.1}$$

We assume that each f_j is symmetric and log-concave. We propose a direct maximum likelihood approach and design a genuine EM algorithm with the usual monotonicity property.

Chang and Walther (2007) have studied a similar mixture model under the assumption that each f_j is log-concave but not necessarily symmetric — obviously, the presence of the location parameter μ_j becomes redundant in that case and the model they consider is really the following model with the assumption of log-concavity,

$$g(x) = \sum_{j=1}^{k} \pi_j f_j(x), \quad \sum_{j=1}^{k} \pi_j = 1, \quad x \in \mathbb{R}. \tag{5.2}$$

The assumption of symmetry is, however, popular, and with that assumption, for each j,

$$f_j^+ := 2 f_j \mathbb{1}_{[0,\infty)} \tag{5.3}$$

is a monotone log-concave density on $[0, \infty)$.

Monotone densities have been used for a variety of applications and their maximum likelihood estimation was first studied by Grenander (1956). Log-concave densities have also been successfully used in nonparametric modeling and their maximum likelihood estimation has been extensively studied in the literature (Balabdaoui, 2004; Balabdaoui et al., 2009; Doss and Wellner, 2016a; Dümbgen and Rufibach, 2009; Rufibach, 2006). However the study of monotone and log-concave densities is what is required to understand the properties of our present model (5.1). In Section 5.1 we prove some basic properties for this class of densities such as uniform consistency of the MLE by simply following the existing literature, and in particular the work of Rufibach (2006).

In Section 5.2 we propose a genuine EM algorithm for fitting the mixture model (5.1). The algorithm includes a step where the monotone and log-concave MLE for f_j^+ is computed. To do so we apply the method[1] of Doss and Wellner (2016b) designed for computing the log-concave MLE with a fixed mode — the mode is of course set to 0 in our case. We note that Balabdaoui and Doss (2014) use the same routine in the numerical implementation of their method.

In Section 5.3 we apply our model to clustering problems and compare our approach with that of (Chang and Walther, 2007) (without symmetry) and that of (Balabdaoui and Doss, 2014) (without a monotone EM and limited to clustering with $k = 2$ components), as well as a Gaussian mixture model (GMM), on both synthetic and real-world datasets, in terms of misclassification errors, Rand Indexes (Rand, 1971), posterior errors, and achieved likelihood. Section 5.4 gathers the contributions as well as the limitations of the study, and summarizes the chapter.

5.1 NPMLE of a monotone and log-concave density

This section is concerned with the estimation of a monotone log-concave density f via maximum likelihood from a given ordered sample $x_1 < x_2 < \cdots < x_n$. We

[1]The method is based on an active set implementation and has been implemented in the R package **logcondens.mode**.

let F denote the distribution function corresponding to the density f and define

$$\psi(x) = \log f(x). \tag{5.4}$$

Requiring that f be monotone and log-concave is equivalent to requiring that ψ is monotone (non-increasing) and concave.

Based on the sample, the negative log-likelihood at f is given by

$$-\sum_{i=1}^{n} \log f(x_i) = -n \sum_{i=1}^{n} \psi(x_i). \tag{5.5}$$

In order to relax the constraint of f being a probability density we follow the trick used by Rufibach (2006) and add a Lagrange term to (5.5), leading to the functional

$$\Lambda_n(\psi) = -\sum_{i=1}^{n} \psi(x_i) + n \int \exp \psi(x) \mathrm{d}x. \tag{5.6}$$

The NPMLE of f is $\hat{f}_n = \exp \hat{\psi}_n$, where $\hat{\psi}_n$ is the minimizer of Λ over class of functions on $[0, \infty)$ that are non-increasing and concave, that is

$$\hat{\psi}_n := \underset{\psi \in \mathcal{MC}}{\arg\min}\, \Lambda_n(\psi), \tag{5.7}$$

where[2]

$$\mathcal{MC} := \big\{\psi : [0, \infty) \to [-\infty, \infty) \mid \psi \text{ is non-increasing, concave, proper, and closed}\big\}. \tag{5.8}$$

The theory below results from a straightforward adaptation of the thesis work of (Rufibach, 2006) on the maximum likelihood of a log-concave density, without the additional constraint of monotonicity, published in the form of a research article in (Dümbgen and Rufibach, 2009). We do not provide proofs but rather refer the reader to that work.

The following results from an adaptation of Theorem 2.1 in (Dümbgen and Rufibach, 2009).

[2]Following the definition in Rockafellar (2015), a concave function f is said to be proper if $f(x) > -\infty$ for at least one x and $f(x) < +\infty$ for every x. A closed function is a function that maps closed sets to closed sets.

Theorem 5 (Existence, uniqueness, and shape)**.** *The NPMLE $\hat{\psi}_n$ exists and is unique. It is linear between sample points and continuous on $[0, x_n]$, with $\hat{\psi}_n(x) = \hat{\psi}_n(x_1)$ for $x \in [0, x_1]$ and $\hat{\psi}_n(x) = -\infty$ for $x > x_n$.*

The following results from an adaptation of Theorem 2.2 in (Dümbgen and Rufibach, 2009).

Theorem 6 (Characterization)**.** *Let ψ be a non-increasing and concave function such that $\{x : \psi(x) > -\infty\} = [0, x_n]$. Then, $\psi = \hat{\psi}_n$ if and only if*

$$\frac{1}{n}\sum_{i=1}^{n}\Delta(x_i) \leq \int \Delta(x)\exp\psi(x)\mathrm{d}x \tag{5.9}$$

for any $\Delta : [0, \infty) \to \mathbb{R}$ such that $\psi + \lambda\Delta$ is non-increasing and concave for some $\lambda > 0$.

For $I \subset \mathbb{R}$ an interval, $\beta \in [1, 2]$, and $L > 0$, let $\mathcal{H}^{\beta,L}(I)$ be the Hölder class of real-valued functions g on I satisfying $|g(y) - g(x)| < L|y - x|$ if $\beta = 1$ and $|g'(y) - g'(x)| \leq L|y - x|^{\beta-1}$ if $\beta \in (1, 2]$, for all $x, y \in I$. The following results from an adaptation of Theorem 4.1 in (Dümbgen and Rufibach, 2009).

Theorem 7 (Uniform consistency)**.** *Assume that $f \in \mathcal{H}^{\beta,L}(I)$ for some exponent $\beta \in [1, 2]$, some constant $L > 0$, and a compact interval $I \subset \{f > 0\}$. Then,*

$$\max_{t\in I}\left|\hat{f}_n(t) - f(t)\right| = O_{\mathbb{P}}(\log n/n)^{\beta/(2\beta+1)}. \tag{5.10}$$

As pointed out by Dümbgen and Rufibach (2009), this is the minimax rate for densities in that smoothness class, as shown by Khas'minskii (1979), so that, when the density is log-concave and Hölder-β (with $\beta \in [1, 2]$) in some interval, the log-concave MLE adapts to the proper smoothness in that interval. We believe the same holds under the additional constraint of monotonicity.

5.2 A Semiparametric EM Algorithm

We now consider fitting the semiparametric mixture model (5.1). Recalling (5.3), this amounts to estimating $\phi := (\boldsymbol{\mu}; \boldsymbol{\pi}; \boldsymbol{f}^+)$ where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k) \in \mathbb{R}^k$,

$\boldsymbol{\pi} = (\pi_1, \ldots, \pi_k)$ is an element of the simplex in $\mathbb{R}^k$, and $\boldsymbol{f}^+ = (f_1^+, \ldots, f_k^+)$ with each f_j^+ being a monotone log-concave density on $[0, \infty)$. Under ϕ, the density of the mixture model is given by

$$g_\phi(x) = \frac{1}{2} \sum_{j=1}^{k} \pi_j f_j^+(|x - \mu_j|). \tag{5.11}$$

The log-likelihood associated of the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ under parameter ϕ is thus given by

$$L(\phi) = \sum_{i=1}^{n} \log g_\phi(x_i). \tag{5.12}$$

It is well-known that directly maximizing $L(\phi)$ is difficult. We design an EM-type algorithm. Let $z_i = j$ when x_i was sampled from the jth component, and define

$$w_{ij} = \mathbb{P}_\phi(z_i = j | x_i) = \frac{\pi_j f_j^+(|x_i - \mu_j|)}{\sum_{l=1}^{k} \pi_l f_l^+(|x_i - \mu_l|)}. \tag{5.13}$$

With these particular weights, clearly,

$$L(\phi) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \log(\pi_j f_j^+(|x_i - \mu_j|)) - C(\boldsymbol{w}), \tag{5.14}$$

where $C(\boldsymbol{w}) := \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \log w_{ij} + n \log 2$. For a set of parameters $\phi = (\boldsymbol{\mu}; \boldsymbol{\pi}; \boldsymbol{f}^+)$ and weights $\boldsymbol{w}_* = (w_{ij*})$, define

$$Q(\phi, \boldsymbol{w}_*) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij*} \log(\pi_j f_j^+(|x_i - \mu_j|)). \tag{5.15}$$

In an iterative implementation, assuming that $\phi_{(t)}$ denotes the set of parameters at iteration t and $\boldsymbol{w}_{(t)}$ the weights computed according to (5.13), a typical EM approach requires the maximization of $Q(\phi, \boldsymbol{w}_{(t)})$ with respect to ϕ. We propose alternative optimization procedure to do so.

The semiparametric EM (SEM) algorithm that we deploy is described below.

- **E-step**: Given $\phi_{(t)}$, we calculate

$$w_{ij(t)} = \mathbb{P}(z_i = j | x_i, \phi_{(t)}) = \frac{\pi_{j(t)} f_{j(t)}^+(|x_i - \mu_{l(t)}|)}{\sum_{l=1}^{k} \pi_{l(t)} f_{l(t)}^+(|x - \mu_{l(t)}|)}, \tag{5.16}$$

  for $i = 1, \ldots, n$ and $j = 1, \ldots, k$.

- **M-step**:

  1. Update $\boldsymbol{\pi}$

$$\pi_{j(t+1)} = \frac{1}{n}\sum_{i=1}^{n} w_{ij(t)}, \quad j = 1, \ldots, k; \tag{5.17}$$

  2. Update $\boldsymbol{\mu}$

$$\mu_{j(t+1)} := \arg\max_{\mu} \sum_{i=1}^{n} w_{ij(t)} \log f^{+}_{j(t)}(|x_i - \mu|), \quad j = 1, \ldots, k; \tag{5.18}$$

  3. Update $\boldsymbol{f}^{+}$

$$f^{+}_{j(t+1)} := \arg\max_{f^{+}} \sum_{i=1}^{n} w_{ij(t)} \log f^{+}(|x_i - \mu_{j(t+1)}|), \quad j = 1, \ldots, k. \tag{5.19}$$

Since $\log f^{+}_{j(t)}$ is concave, the objective function in (5.18) is a concave function of μ, therefore Golden Section Search can be applied to solve this optimization problem. In (5.19), the optimization is over f^{+} being a monotone and log-concave density on $[0, \infty)$. The solution corresponds to the weighted NPMLE based on data $(|x_1 - \mu_{j(t+1)}|, \ldots, |x_n - \mu_{j(t+1)}|)$ and weights $(w_{1j(t)}, \ldots, w_{nj(t)})$.

Remark 3. Our implementation is based on applying the function activeSetLogCon.mode in the R package logcondens.mode with mode chosen to be 0.

- **Initialization**: We initialize $\boldsymbol{w}_{(0)}$ and $\boldsymbol{f}^{+}_{(0)}$ at the values given by a fit of a GMM, and start with M-step first.

Our SEM algorithm has the desirable monotonicity property of a true EM algorithm (Dempster et al., 1977; Wu, 1983).

Proposition 1 (Monotonicity property)**.** *With the same notation,* $L(\phi_{(t)}) \leq L(\phi_{(t+1)})$ *for all* $t \geq 0$.

Proof. In the algorithm, armed with $\phi_{(t)}$, we compute the weights $\boldsymbol{w}_{(t)}$ in the E-step and in the M-step we obtain $\phi_{(t+1)}$ by maximizing $Q(\phi, \boldsymbol{w}_{(t)})$ over ϕ. (We do the latter sequentially, first over $\boldsymbol{\pi}$, then over $\boldsymbol{\mu}$, and finally over $\boldsymbol{f}^{+}$.) In particular,

$$Q(\phi_{(t+1)}, \boldsymbol{w}_{(t)}) \geq Q(\phi_{(t)}, \boldsymbol{w}_{(t)}). \tag{5.20}$$

The key, then, is Jensen's inequality, which implies that for a set of parameters $\phi = (\boldsymbol{\mu}; \boldsymbol{\pi}; \boldsymbol{f}^+)$ and non-negative weights $\boldsymbol{w}_* = (w_{ij*})$ such that $\sum_j w_{ij*} = 1$ for all i,

$$\begin{aligned} L(\phi) &= \sum_{i=1}^{n} \log\left(\frac{1}{2}\sum_{j=1}^{k} \pi_j f_j^+(|x_i - \mu_j|)\right) && (5.21) \\ &= \sum_{i=1}^{n} \log\left(\sum_{j=1}^{k} w_{ij*} \frac{\pi_j f_j^+(|x_i - \mu_j|)}{w_{ij*}}\right) - n\log 2 \\ &\geq \sum_{i=1}^{n}\sum_{j=1}^{k} w_{ij*} \log\left(\frac{\pi_j f_j^+(|x_i - \mu_j|)}{w_{ij*}}\right) - n\log 2 \\ &= Q(\phi, \boldsymbol{w}_*) - C(\boldsymbol{w}_*), && (5.22) \end{aligned}$$

with equality if the weights $\boldsymbol{w}_*$ are the weights associated with ϕ as specified in (5.13). In particular,

$$L(\phi_{(t+1)}) \geq Q(\phi_{(t+1)}, \boldsymbol{w}_{(t)}) - C(\boldsymbol{w}_{(t)}), \tag{5.23}$$

while

$$L(\phi_{(t)}) = Q(\phi_{(t)}, \boldsymbol{w}_{(t)}) - C(\boldsymbol{w}_{(t)}). \tag{5.24}$$

We thus have

$$\begin{aligned} L(\phi_{(t+1)}) &\geq Q(\phi_{(t+1)}, \boldsymbol{w}_{(t)}) - C(\boldsymbol{w}_{(t)}) \\ &\geq Q(\phi_{(t)}, \boldsymbol{w}_{(t)}) - C(\boldsymbol{w}_{(t)}) = L(\phi_{(t)}). \end{aligned}$$ □

5.3 Numerical experiments

We now consider the problem of one-dimensional clustering. We assume that the data can be clustered into k groups, fit the k-component mixture (5.1) as described in Section 5.2 obtaining $\hat{\phi}$, and assign a label to an observation x_i according to Bayes optimal rule

$$\arg\max_j \ \mathbb{P}(z_i = j | x_i, \hat{\phi}) = \arg\max_j \ \frac{\hat{\pi}_j \hat{f}_j^+(|x_i - \hat{\mu}_j|)}{\sum_{j=1}^{k} \hat{\pi}_l \hat{f}_l^+(|x_i - \hat{\mu}_l|)}. \tag{5.25}$$

We apply our SEM algorithm both on simulated and real data. In Section 5.3.1, we choose to simulate data from the Gaussian and Laplace mixture models used in (Balabdaoui and Doss, 2014), and in Section 5.3.2, we apply the SEM algorithm to the well-known Old Faithful geyser data also investigated in (Balabdaoui and Doss, 2014), and to the rainfall data studied in (Bordes et al., 2006).

5.3.1 Synthetic datasets

As a first example, we use a two-component Gaussian mixture to empirically check the convergence of our SEM algorithm. We first sample $n = 100$ (Figure 5.1) and then sample $n = 300$ (Figure 5.2) observations from the Gaussian mixture $0.15\,\mathcal{N}(-1, 1) + 0.85\,\mathcal{N}(2, 1)$ and apply the SEM algorithm to these two datasets respectively. This seems to be the most difficult situation considered in (Bordes et al., 2006). Panels (a), (b), (c), and (d) of Figure 5.1 show that SEM stabilizes after about 8 iterations for the three Euclidean parameters and the observed data likelihood. As expected, the achieved maximum data likelihood is monotonically increasing as a function of the number of iterations. Panels (e) and (f) show the final NPMLE for f_1, f_2, g and compare that with the truth. The NPMLE for the symmetric log-concave densities are piecewise exponential, which is consistent with what is described in Theorem 5. Figure 5.2 is provided to show on the improvement resulting from a larger sample size. With initialization manually set the same with that in Figure 5.1, the effect is visible on the recovering of f_1^+, f_2^+ and g.

We then conduct a Monte Carlo study to compare the performance of our algorithm (SEM) in clustering with the methods proposed in (Chang and Walther, 2007) and (Balabdaoui and Doss, 2014). Chang and Walther fit the simple mixture model (5.2) and only assume that the components are log-concave densities. Balabdaoui and Doss fit the semiparametric model (2.23) and employ the parameter estimators from (Hunter et al., 2007), and then assumes that both components have the same density after centering and fits that density using the symmetric log-concave density estimator. We denote these two methods by LCM and SLC respectively. We compare SEM, LCM and SLC to GMM, which serves as benchmark when the underlying model is indeed in that class. We compare these

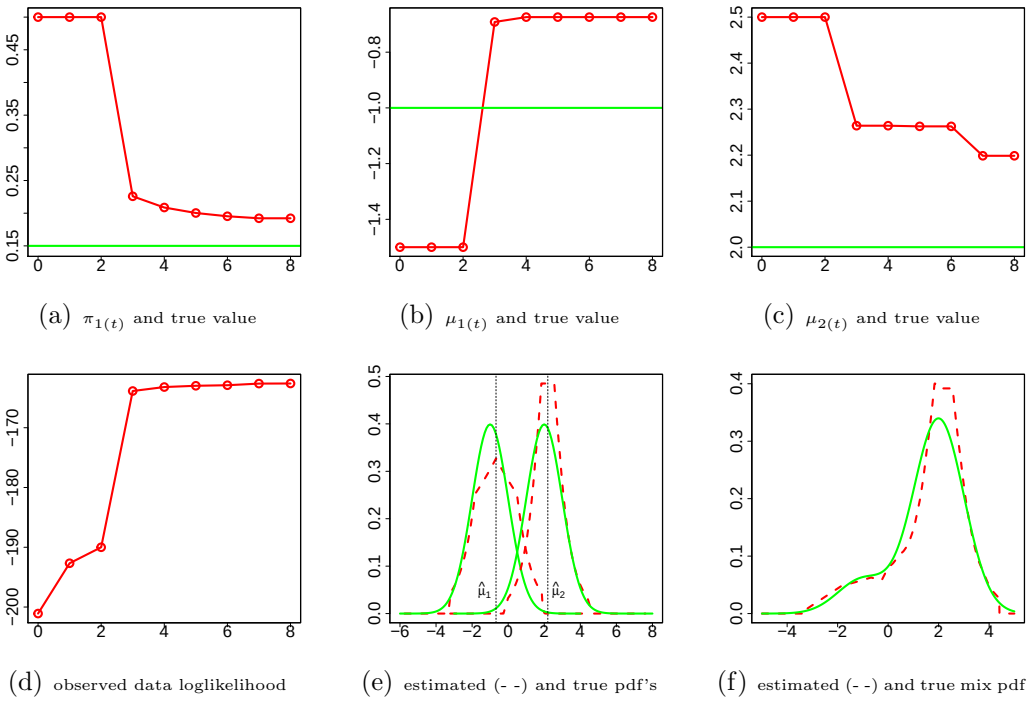
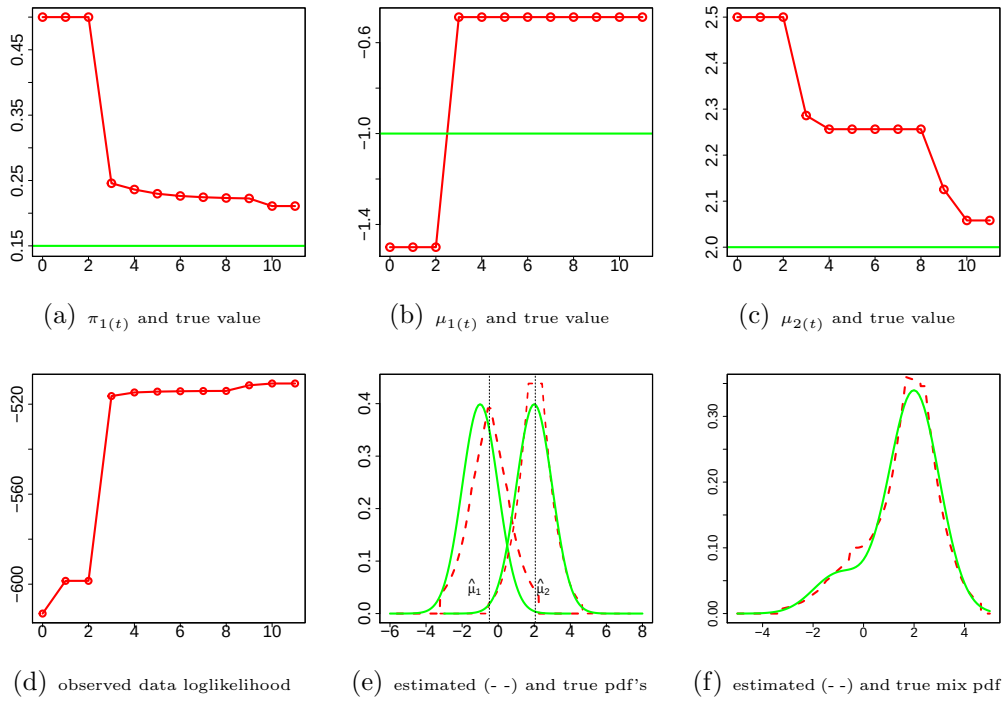


Figure 5.1: SEM for the Gaussian mixture with $n = 100, \pi_1 = 0.15, \mu_1 = -1$ and $\mu_2 = 2$.

methods on two Gaussian mixture models, two Laplace mixture models and one Gaussian-Laplace mixture model described below:

- Model 1: $0.2\,\mathcal{N}(0,1) + 0.8\,\mathcal{N}(1,1)$;
- Model 2: $0.2\,\mathcal{N}(0,1) + 0.8\,\mathcal{N}(2,2)$;
- Model 3: $0.2\,\mathcal{L}(0,1) + 0.8\,\mathcal{L}(1,1)$;
- Model 4: $0.2\,\mathcal{L}(0,1) + 0.4\,\mathcal{L}(1.5,1) + 0.4\,\mathcal{L}(-1.5,1)$;
- Model 5: $0.2\,\mathcal{N}(0,1)+0.2\,\mathcal{N}(1.5,1)+0.2\,\mathcal{N}(-1.5,1)+0.2\,\mathcal{L}(3,1)+0.2\,\mathcal{L}(-3,1)$.

The sample size is $n = 500$ for the first 4 models and $n = 1000$ for Model 5. Each setting is repeated 1000 times. We examine the quality of the resulting clustering in terms of the achieved data log-likelihood, the misclassification errors when $k = 2$ or Rand Indexes when $k \geq 3$, and the average absolute posterior probability error used by (Chang and Walther, 2007) — all averaged over the 1000



(a) $\pi_{1(t)}$ and true value (b) $\mu_{1(t)}$ and true value (c) $\mu_{2(t)}$ and true value

(d) observed data loglikelihood (e) estimated (- -) and true pdf's (f) estimated (- -) and true mix pdf

Figure 5.2: SEM for the Gaussian mixture, $n = 300, \pi_1 = 0.15, \mu_1 = -1$ and $\mu_2 = 2$.

repeats. The latter metric investigates how well a mixture clustering algorithm estimates the uncertainty for the membership assignment of each observation on population level. This metric is defined as

$$\text{posterior error} := \frac{1}{n}\sum_{i=1}^{n} |\hat{w}_{i1} - w_{i1}|, \tag{5.26}$$

where $\hat{w}_{i1}$ and w_{i1} are computed by (5.13) with estimators and true parameters respectively. Notice that this metric only applies to clustering with $k = 2$ components. When $k \geq 3$, we define the posterior error by the Frobenius distance

$$\text{posterior error} := \min_{P_\pi}\{\|\hat{w}P_\pi - w\|_F\}, \tag{5.27}$$

where P_π is any $k \times k$ permutation matrix, and matrices $\hat{w}$ and w are computed by (5.13) with estimators and true parameters respectively. We report the comparison results in Table 5.1. As can be seen from this table, GMM, LCM, and our SEM algorithm clearly outperform SLC in terms of log-likelihood and posterior error.

Table 5.1: Comparison of the four different clustering methods in terms of achieved log-likelihood, number of misclassification errors (when $k = 2$) or Rand index (when $k > 2$), and posterior errors. The reported numbers are the average of the metrics over $R = 1000$ replications under each of the three symmetric and log-concave mixture models. Each time the sample size is $n = 500$ for Model $1 \sim 4$ and $n = 1000$ for Model 5. The numbers in parentheses are the corresponding standard errors.

	Metric	GMM	LCM	SLC	SEM
Model 1	log-like	-743.2 (0.50)	–739.4 (0.51)	-1104.9 (2.06)	**-738.6 (0.50)**
	mis-class	122.7 (1.31)	**102.6 (1.49)**	174.2 (1.22)	123.7 (1.30)
	post-error	**0.199 (0.003)**	0.206 (0.002)	0.317 (0.001)	0.202 (0.003)
Model 2	log-like	-1049.7 (0.48)	-1046.3 (0.49)	-1383.0 (2.43)	**-1044.7 (0.48)**
	mis-class	148.1 (1.55)	125.4 (1.91)	**118.4 (0.70)**	150.0 (1.54)
	post-error	**0.211 (0.004)**	0.255 (0.003)	0.216 (0.004)	0.283 (0.001)
Model 3	log-like	-876.0 (0.67)	**-869.3 (0.69)**	-1293.6 (3.94)	-870.6 (0.66)
	mis-class	162.8 (1.12)	**111.4 (1.23)**	153.2 (1.09)	159.1 (1.18)
	post-error	0.236 (0.002)	0.244 (0.002)	0.324 (0.001)	**0.234 (0.002)**
Model 4	log-like	-1031.1 (16.8)	-1030.0 (16.6)	-	**-1025.7 (16.8)**
	Rand index	0.602 (0.097)	**0.655 (0.118)**	-	0.607 (0.093)
	post-error	**10.6 (2.65)**	13.3 (1.53)	-	**10.6 (2.61)**
Model 5	log-like	-2281.2 (19.9)	-2281.8 (19.8)	-	**-2275.7 (19.5)**
	Rand index	0.622 (0.138)	0.391(0.235)	-	**0.623 (0.137)**
	post-error	**18.4 (3.15)**	20.1 (3.32)	-	18.6 (3.23)

LCM outperforms other methods in terms of misclassification error or Rand index when $k \leq 3$, but does not perform well when $k = 5$. When the mixture densities are normal, SEM performs as well as GMM, arguably the gold standard in such a situation; when the densities are Laplace, SEM slightly improves the clustering initialized by GMM. Moreover, SEM achieves a significantly higher log-likelihood compared with the other methods when the mixture densities are normal. We also notice that SLC sometimes gives better results in terms of misclassification error, even though the posterior-error is worse.

5.3.2 Real datasets

In this section, we apply our new estimation approach to two different real-world datasets. Both of these datasets are included in the standard R distribution.

The first dataset consists of times, in minutes, between eruptions of the Old Faithful geyser in Yellowstone National Park. Figure 5.3 plots the iterations of our SEM algorithm, which is seen to converge rather quickly, in less than 14 iterations. Table 5.2 shows that our estimates are close to those obtained by GMM, Hunter et al. (2007), and Bordes et al. (2007), while the estimates from Balabdaoui and Doss (2014) are a bit farther away.

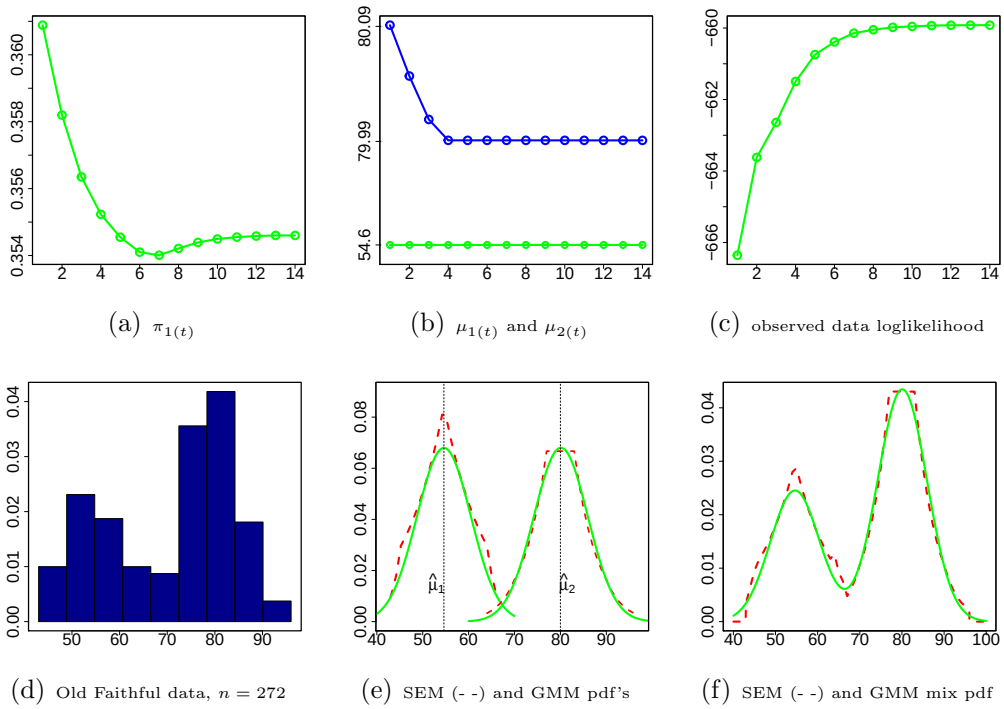


Figure 5.3: SEM applied to the Old Faithful waiting data.

The second dataset is the average amount of precipitation (rainfall) in inches for each of 70 cities in the United States and Puerto Rico (McNeil, 1977). Figure 5.4 plots the iterations of our SEM algorithm, which is again seen to converge quickly.

Table 5.2: Parameter estimates for the Old Faithful geyser waiting data, using GMM, the semiparametric estimation from Hunter et al. (2007)(SP), the stochastic EM algorithm by Bordes et al. (2007) (SP-EM), the symmetric log-concave mixture model by Balabdaoui and Doss (2014) (SLC) and our SEM algorithm.

parameters	GMM	SP	SP-EM	SLC	SEM
π_1	0.361	0.352	0.359	0.33	0.355
μ_1	54.61	54.0	54.59	55.5	54.61
μ_2	80.09	80.0	80.05	80.5	80.5

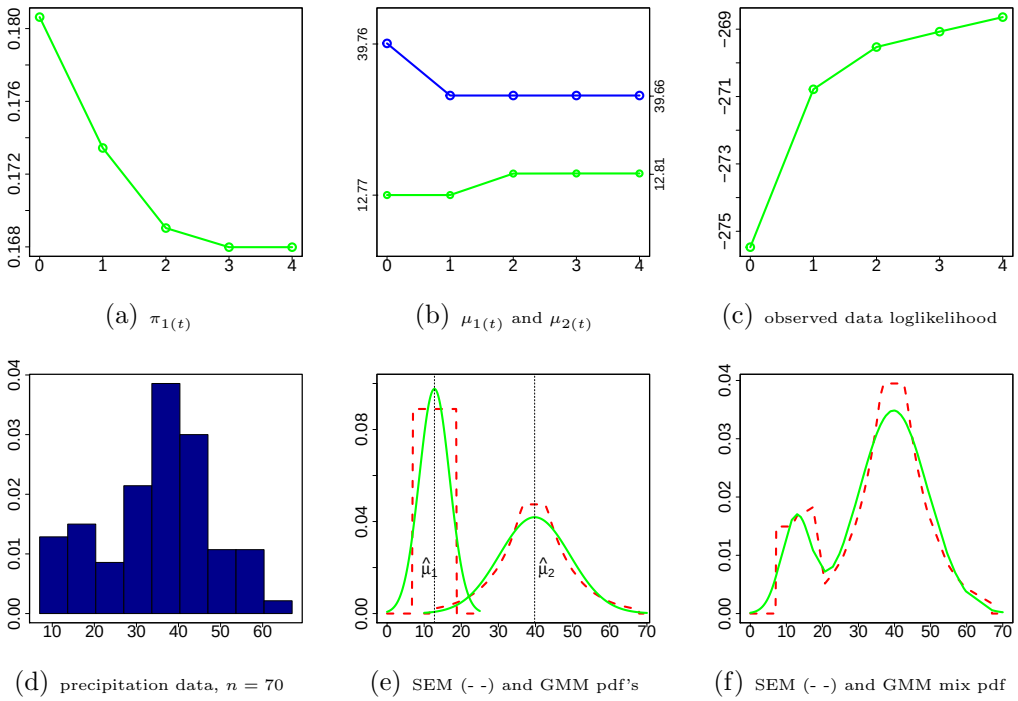
5.4 Discussion

In this chapter, we revisited the problem of fitting a mixture model under the assumption that the mixture components are symmetric and log-concave. We studied the nonparametric MLE of a monotone and log-concave probability density and provided some basic properties for this class of densities such as uniform consistency. We then developed a semiprametric EM algorithm which possess the monotone property of a genuine EM algorithm and can be adapted to other semiparametric mixture models. Numerical studies on both synthetic datasets and real-world datasets, show that our algorithm improves on the method of Balabdaoui and Doss (2014).

Our study in this chapter only considered univariate semiparametric mixtures. Since the computation time of the MLE of multivariate log-concave densities becomes quickly intractable as the dimension increases, extending log-concave mixture models to higher dimensions presents a real challenge. Chang and Walther (2007) propose using a normal copula and perform simulations in dimension two. Multivariate nonparametric mixture models with KDE are developed in (Benaglia et al., 2009; Chauveau and Hoang, 2016; Chauveau et al., 2015; Levine et al., 2011).

Acknowledgement

This chapter, in full, has been organized into the following paper: *Semiparametric Estimation of Symmetric Mixture Models with Monotone and Log-Concave*



(a) $\pi_{1(t)}$ (b) $\mu_{1(t)}$ and $\mu_{2(t)}$ (c) observed data loglikelihood

(d) precipitation data, $n = 70$ (e) SEM (- -) and GMM pdf's (f) SEM (- -) and GMM mix pdf

Figure 5.4: SEM applied to the annual precipitation data.

Densities (Xiao Pu and Ery Arias-Castro), and has been submitted for publication. The dissertation author is the primary investigator and corresponding author of this material.

Chapter 6

Concentration of Measure for Radial Distribution

When fitting multivariate mixture models, it is quite tempting to extend the Gaussian mixture models to models of the form

$$\sum_{k=1}^{K} \pi_k |A_k| g_k(\|A_k x\|), \tag{6.1}$$

where we assume the mixture has K components, with the kth component having weight π_k and density $|A_k| g_k(\|A_k x\|)$. For example, Bickel et al. (1998) and more recently Bhattacharyya and Bickel (2015) consider models of this kind. Instead of smoothness assumptions, we are more interested here in shape assumptions, for example that g_k is decreasing and/or log-concave on $\mathbb{R}_+$. Chang and Walther (2007) consider such mixture models but under the Naive Bayes assumption instead of assuming the densities are elliptical. An EM approach to fitting such a model involves being able to estimate g_k based on a sample from $g_k(\|x\|)$. And this is what we found challenging in our investigation.

Focusing on this task, suppose we have an i.i.d. sample from f, where f rotationally invariant (aka radial), meaning that $f(x) = g(\|x\|)$ for some function g, and consider the problem of estimating g. In fact, we can work with the magnitudes (the norms of the observations), which are sufficient. We explain the difficulty of estimating g by the fact that the magnitudes are highly concentrated as the dimension becomes large.

The simplest case of this concentration of measure phenomenon arises when we assume that g is proportional to ψ, where ψ is fixed, as is the case in the Gaussian setting. Specifically, we assume we are in dimension $d+1$ and we work with ψ satisfying the following assumptions (and some additional assumptions specified later on)

$$\psi : \mathbb{R}_+ \to \mathbb{R}_+ \text{ is such that } 1/c_d := \int_0^\infty u^d \psi(u) \mathrm{d}u < \infty \text{ for all } d \geq 1. \tag{6.2}$$

We let X_d denote a random variable with density $c_d\psi(\|x\|)$ on $\mathbb{R}^{d+1}$ and let U_d denote its magnitude, $U_d = \|X_d\|$, which has density $c_d u^d \psi(u)$ on $\mathbb{R}_+$.

In this context we show a form of concentration of measure, and convergence in distribution, as the dimension d increases. Concentration is a well-known phenomenon in high-dimensions, in particular for product distributions (Naive Bayes), with far-reaching consequences (Boucheron et al., 2013; Ledoux, 2005). For radial distributions, it is not as well-known, except for when the density is Gaussian or uniform on a ball. (The latter is often used to explain some forms of curse of dimensionality.)

In Section 6.1 we study the case where ψ has compact support, which is the simplest situation. In Section 6.2 we consider the case where ψ is *not* compactly supported. In Section 6.3 we discuss our results and some possible implications for statistical modeling.

6.1 The case of compact support

In this whole section we assume that ψ has compact support. Define the supremum of the support as follows

$$u_* = \sup\Big\{u : \int_{u-\varepsilon}^{u} \psi(u)\mathrm{d}u > 0 \text{ for all } \varepsilon > 0\Big\}. \tag{6.3}$$

Note that $u_* < \infty$ by assumption and that the support of ψ is included in $[0, u_*]$. If ψ is continuous (which the reader can assume without much loss of generality), then the following is an equivalent definition $u_* = \sup\{u : \psi(u) > 0\}$. The emblematic example is that of the uniform distribution on the unit ball, in which

case $\psi(u) = \mathbb{I}\{u \leq 1\}$ and $u_* = 1$. This distribution is well-known to concentrate near the boundary of its support (the unit sphere). Our results below extend this to other distributions with compact support.

6.1.1 Convergence in probability

We start by establishing a convergence in probability.

Theorem 8. *In the setting considered here, $U_d \to u_*$ in probability as $d \to \infty$.*

Proof. Assume $u_* = 1$ without loss of generality. Then

$$\mathbb{P}(U_d < 1 - \varepsilon) = c_d \int_0^{1-\varepsilon} u^d \psi(u) \mathrm{d}u \leq c_d (1-\varepsilon)^d \int_0^1 \psi(u) \mathrm{d}u, \tag{6.4}$$

while

$$\mathbb{P}(U_d \geq 1-\varepsilon) \geq \mathbb{P}(U_d \geq 1-\varepsilon/2) = c_d \int_{1-\varepsilon/2}^1 u^d \psi(u) \mathrm{d}u \geq c_d (1-\varepsilon/2)^d \int_{1-\varepsilon/2}^1 \psi(u) \mathrm{d}u. \tag{6.5}$$

Note that the last integral is strictly positive for all $\varepsilon > 0$ by definition of u_* in (6.3) (recall that we assumed that $u_* = 1$). Hence

$$\frac{\mathbb{P}(U_d < 1 - \varepsilon)}{\mathbb{P}(U_d \geq 1 - \varepsilon)} \leq \frac{(1-\varepsilon)^d \int_0^1 \psi(u) \mathrm{d}u}{(1-\varepsilon/2)^d \int_{1-\varepsilon/2}^1 \psi(u) \mathrm{d}u} \to 0, \quad d \to \infty, \tag{6.6}$$

when $\varepsilon \in (0, 1)$ is fixed. Since $\mathbb{P}(U_d < 1 - \varepsilon) + \mathbb{P}(U_d \geq 1 - \varepsilon) = 1$, we proved that $\mathbb{P}(U_d < 1 - \varepsilon) \to 0$ for all $\varepsilon > 0$. This, coupled with the fact that $\mathbb{P}(U_d \leq 1) = 1$, proves that $U_d \to 1$ in probability as $d \to \infty$. □

6.1.2 Convergence in distribution

Beyond a convergence in probability, we can establish a convergence in distribution. The limiting distribution happens to depend on the behavior of ψ in the neighborhood of u_*. We only cover the case where ψ behaves as a power function near u_*.

Theorem 9. *In the setting considered here, assume in addition that ψ is bounded and that $\psi(u) \sim a(u_* - u)^b$ as $u \nearrow u_*$ for some $a > 0$ and $b > -1$. Then $d(u_* - U_d)$*

converges weakly to the Gamma distribution with shape parameter $b+1$ and rate $1/u_*$.

Proof. Assume without loss of generality that $u_* = 1$. We first control the behavior of c_d as $d \to \infty$. Fix $\varepsilon \in (0,1)$. On the one hand, by the assumptions on ψ and Dominated Convergence, we have

$$\int_{1-\varepsilon}^{1} u^d \psi(u) \mathrm{d}u \sim \int_{1-\varepsilon}^{1} u^d a(1-u)^b \mathrm{d}u \sim aB(d+1, b+1), \quad d \to \infty, \tag{6.7}$$

where B is the Beta function. On the other hand, by Theorem 8,

$$c_d \int_{1-\varepsilon}^{1} u^d \psi(u) \mathrm{d}u \sim 1, \quad d \to \infty. \tag{6.8}$$

Together, this proves that

$$1/c_d \sim aB(d+1, b+1) \sim a\Gamma(b+1) d^{-(b+1)}, \quad d \to \infty, \tag{6.9}$$

where Γ is the Gamma function.

We now consider the case where $\varepsilon = \varepsilon_d \to 0$ as $d \to \infty$. More precisely, we fix $t > 0$ and set $\varepsilon_d = t/d$. By Dominated Convergence again, applied twice, and a change of variables, as $d \to \infty$, we have

$$\begin{aligned}
\mathbb{P}(U_d > 1 - t/d) &= c_d \int_{1-t/d}^{1} u^d \psi(u) \mathrm{d}u && (6.10)\\
&\sim c_d \int_{1-t/d}^{1} u^d a(1-u)^b \mathrm{d}u && (6.11)\\
&= c_d a d^{-(b+1)} \int_0^t (1 - v/d)^d v^b \mathrm{d}v && (6.12)\\
&\sim \frac{1}{\Gamma(b+1)} \int_0^t e^{-v} v^b \mathrm{d}v. && (6.13)
\end{aligned}$$

Recognizing the distribution function of the Gamma distribution with shape parameter $b+1$ and rate 1, the proof is complete. □

6.2 The case of non-compact support

We now assume that ψ has non-compact support, which is equivalent to $u_* = \infty$ in (6.3). We note that here the emblematic example is that of the standard normal

distribution, which is known to concentrate near the sphere of radius $\sqrt{d}$, meaning $U_d/\sqrt{d} \to 1$. In fact, U_d^2 has the chi-squared distribution with d degrees of freedom, and in particular, $\sqrt{2}(U_d - \sqrt{d})$ is asymptotically standard normal in the limit $d \to \infty$. Our results below extend this phenomena to other radial distributions.

While we were able to handle the case of compact support, which we treated in Section 6.1, with very natural assumptions, the case of non-compact support appears more challenging and our working assumptions are more complicated. This is despite the fact that we favored simplicity over generality. Nevertheless, our working assumptions include interesting (and natural) examples.

6.2.1 Convergence in probability

We start by establishing a convergence in probability.

We start by making the following assumptions. We assume there is $u_\ddagger$ such that, for $u \geq u_\ddagger$, $\Lambda(u) := -\log \psi(u)$ is differentiable and $L(u) := u\Lambda'(u)$ is increasing. In addition, we assume that $M(u) := L(u)/\log(u) \to \infty$ as $u \to \infty$ and

$$\limsup_{u\to\infty} \frac{M((1-\varepsilon)u)}{M(u)} \leq 1, \quad \liminf_{u\to\infty} \frac{M((1+\varepsilon)u)}{M(u)} \geq 1, \quad \forall \varepsilon \in (0,1). \tag{6.14}$$

Theorem 10. *In the setting considered here, $U_d/u_d \to 1$ in probability as $d \to \infty$, where $u_d := L^{-1}(d)$.*

Example 1. Consider the case where $\Lambda(u) = c\log(u+a)^\alpha(u+b)^\beta$, where $a > 0$, $b \geq 0$, $c > 0$, $\alpha \in \mathbb{R}$ and $\beta > 0$. Surely, this defines a bonafide shape function ψ in the sense of (6.2). It can be shown that ψ defined as such satisfies the conditions of Theorem 10, with

$$u_d \sim c^{-1/\beta}\beta^{(\alpha-1)/\beta}(\log d)^{-\alpha/\beta}d^{1/\beta}, \quad d \to \infty. \tag{6.15}$$

Proof. The function $u \mapsto u^d\psi(u)$ is increasing on $[0, u_d)$ and decreasing on (u_d, ∞). Indeed, $\log(u^d\psi(u)) = d\log u - \Lambda(u)$ has derivative $\frac{1}{u}(d - L(u))$, which is positive for $u < u_d$, zero at $u = u_d$, and negative at $u > u_d$, by our assumptions and the definition of u_d. Note that, necessarily, $u_d \to \infty$ as $d \to \infty$.

We have

$$\mathbb{P}(U_d \leq v) = c_d \int_0^v u^d\psi(u)\mathrm{d}u. \tag{6.16}$$

Fix $\varepsilon \in (0,1)$.

Left tail. Using the fact that $u^d\psi(u) \le u_0^d\psi(u_0)$ for any $u \le u_0 \le u_d$, we have

$$\frac{1}{c_d}\,\mathbb{P}(U_d \le (1-\varepsilon)u_d) = \int_0^{(1-\varepsilon)u_d} u^d\psi(u)\mathrm{d}u \tag{6.17}$$

$$\le ((1-\varepsilon)u_d)^{d+1}\psi((1-\varepsilon)u_d), \tag{6.18}$$

and we also have

$$\frac{1}{c_d}\,\mathbb{P}(U_d \ge (1-\varepsilon)u_d) = \int_{(1-\varepsilon)u_d}^{\infty} u^d\psi(u)\mathrm{d}u \tag{6.19}$$

$$\ge \int_{(1-\varepsilon/2)u_d}^{u_d} u^d\psi(u)\mathrm{d}u \tag{6.20}$$

$$\ge (\varepsilon/2)u_d((1-\varepsilon/2)u_d)^d\psi((1-\varepsilon/2)u_d). \tag{6.21}$$

Taking the ratio, we obtain

$$\frac{\mathbb{P}(U_d \le (1-\varepsilon)u_d)}{\mathbb{P}(U_d \ge (1-\varepsilon)u_d)} \le \frac{((1-\varepsilon)u_d)^{d+1}\psi((1-\varepsilon)u_d)}{(\varepsilon/2)u_d((1-\varepsilon/2)u_d)^d\psi((1-\varepsilon/2)u_d)} \tag{6.22}$$

$$\le \frac{1-\varepsilon}{\varepsilon/2}\left(\frac{1-\varepsilon}{1-\varepsilon/2}\right)^d \frac{\psi((1-\varepsilon)u_d)}{\psi((1-\varepsilon/2)u_d)}. \tag{6.23}$$

Applying the logarithm, and ignoring the constant factor, we further get

$$d\log\left(\frac{1-\varepsilon}{1-\varepsilon/2}\right) - \Lambda((1-\varepsilon)u_d) + \Lambda((1-\varepsilon/2)u_d) \tag{6.24}$$

$$= -d\int_{(1-\varepsilon)u_d}^{(1-\varepsilon/2)u_d} \frac{1}{u}\mathrm{d}u + \int_{(1-\varepsilon)u_d}^{(1-\varepsilon/2)u_d} \Lambda'(u)\mathrm{d}u \tag{6.25}$$

$$= -\int_{(1-\varepsilon)u_d}^{(1-\varepsilon/2)u_d} (d - L(u))\frac{1}{u}\mathrm{d}u \tag{6.26}$$

$$\le -(d - L((1-\varepsilon/2)u_d))\log\left(\frac{1-\varepsilon/2}{1-\varepsilon}\right), \tag{6.27}$$

where we used the monotonicity of L in the last line. Therefore, to show that the fraction in (6.22) converges to 0, it suffices to show that $d - L((1-\varepsilon/2)u_d) \to \infty$. The limit is as $d \to \infty$ while ε remains fixed. Using the fact that $L(u_d) =$

$M(u_d)\log u_d = d$, we have

$$d - L((1-\varepsilon/2)u_d) = d - M((1-\varepsilon/2)u_d)\log((1-\varepsilon/2)u_d) \tag{6.28}$$

$$= d - M((1-\varepsilon/2)u_d)\log(1-\varepsilon/2) - M((1-\varepsilon/2)u_d)\frac{d}{M(u_d)} \tag{6.29}$$

$$= -M((1-\varepsilon/2)u_d)\log(1-\varepsilon/2) + d\left[1 - \frac{M((1-\varepsilon/2)u_d)}{M(u_d)}\right]. \tag{6.30}$$

In the last line, the first term tends to infinity because $M(u) \to \infty$ as $u \to \infty$ and $u_d \to \infty$, while the second terms is nonnegative in the limit because of (6.14), so that the last expression tends to infinity.

We conclude that, for the left tail,

$$\mathbb{P}(U_d \le (1-\varepsilon)u_d) \to 0, \quad d \to \infty. \tag{6.31}$$

Right tail. Using the fact that $u^\ell\psi(u) \le u_0^\ell\psi(u_0)$ for any $u \ge u_0 \ge u_\ell$, where $u_\ell = L^{-1}(\ell)$ in congruence with our definition above, and assuming for now that $b := L((1+\varepsilon)u_d) > d+1$, we have

$$\frac{1}{c_d}\,\mathbb{P}(U_d \ge (1+\varepsilon)u_d) = \int_{(1+\varepsilon)u_d}^{\infty} u^d\psi(u)\mathrm{d}u \tag{6.32}$$

$$\le ((1+\varepsilon)u_d)^b\psi((1+\varepsilon)u_d)\int_{(1+\varepsilon)u_d}^{\infty} u^{d-b}\mathrm{d}u \tag{6.33}$$

$$= ((1+\varepsilon)u_d)^b\psi((1+\varepsilon)u_d)\frac{((1+\varepsilon)u_d)^{d+1-b}}{b-d-1} \tag{6.34}$$

$$= \frac{((1+\varepsilon)u_d)^{d+1}}{b-d-1}\psi((1+\varepsilon)u_d), \tag{6.35}$$

and we also have

$$\frac{1}{c_d}\,\mathbb{P}(U_d \le (1+\varepsilon)u_d) = \int_0^{(1+\varepsilon)u_d} u^d\psi(u)\mathrm{d}u \tag{6.36}$$

$$\ge \int_{u_d}^{(1+\varepsilon/2)u_d} u^d\psi(u)\mathrm{d}u \tag{6.37}$$

$$\ge (\varepsilon/2)u_d((1+\varepsilon/2)u_d)^d\psi((1+\varepsilon/2)u_d). \tag{6.38}$$

Taking the ratio, we obtain

$$\frac{\mathbb{P}(U_d \geq (1+\varepsilon)u_d)}{\mathbb{P}(U_d \leq (1+\varepsilon)u_d)} \leq \frac{\frac{((1+\varepsilon)u_d)^{d+1}}{b-d-1}\psi((1+\varepsilon)u_d)}{(\varepsilon/2)u_d((1+\varepsilon/2)u_d)^d\psi((1+\varepsilon/2)u_d)} \tag{6.39}$$

$$\leq \frac{1+\varepsilon}{\varepsilon/2}\frac{1}{b-d-1}\left(\frac{1+\varepsilon}{1+\varepsilon/2}\right)^d \frac{\psi((1+\varepsilon)u_d)}{\psi((1+\varepsilon/2)u_d)}. \tag{6.40}$$

We pause to show that $b-d \to \infty$ eventually. This is because, using the fact that $L(u_d) = M(u_d)\log u_d = d$,

$$b - d = L((1+\varepsilon)u_d) - d = M((1+\varepsilon)u_d)\log((1+\varepsilon)u_d) - d \tag{6.41}$$

$$= M((1+\varepsilon)u_d)\log(1+\varepsilon) + \left[\frac{M((1+\varepsilon)u_d)}{M(u_d)} - 1\right] d. \tag{6.42}$$

In the last line, the first term tends to infinity because $M(u) \to \infty$ as $u \to \infty$ and $u_d \to \infty$, while the second terms is nonnegative in the limit because of (6.14), so that the last expression tends to infinity.

Returning to (6.39), applying the logarithm, and ignoring the first two factors whose product is bounded by 1 eventually, we further get

$$d\log\left(\frac{1+\varepsilon}{1+\varepsilon/2}\right) - \Lambda((1+\varepsilon)u_d) + \Lambda((1+\varepsilon/2)u_d) \tag{6.43}$$

$$= d\int_{(1+\varepsilon/2)u_d}^{(1+\varepsilon)u_d} \frac{1}{u}\mathrm{d}u - \int_{(1+\varepsilon/2)u_d}^{(1+\varepsilon)u_d} \Lambda'(u)\mathrm{d}u \tag{6.44}$$

$$= -\int_{(1+\varepsilon/2)u_d}^{(1+\varepsilon)u_d} (L(u)-d)\frac{1}{u}\mathrm{d}u \tag{6.45}$$

$$\leq -(L((1+\varepsilon/2)u_d) - d)\log\left(\frac{1+\varepsilon}{1+\varepsilon/2}\right), \tag{6.46}$$

where we used the monotonicity of L in the last line. Therefore, to show that the fraction in (6.39) converges to 0, it suffices to show that $L((1+\varepsilon/2)u_d) - d \to \infty$, and we already did this in (6.41).

We conclude that, for the right tail,

$$\mathbb{P}(U_d \geq (1+\varepsilon)u_d) \to 0, \quad d \to \infty. \tag{6.47}$$

We can therefore conclude that $U_d/u_d \to 1$ in probability as $d \to \infty$. □

6.2.2 Convergence in distribution

We now turn to establishing a convergence in distribution. Although we speculate that other cases may arise, we give (additional) sufficient conditions for a Gaussian limit.

We make the following additional assumptions. We assume that L is differentiable with $\nu_d := u_d L'(u_d) \to \infty$ and that there is $\omega_d \to \infty$ such that

$$\left|L(u_d) - L((1-\varepsilon)u_d) - \varepsilon u_d L'(u_d)\right| \le |\varepsilon|\nu_d/\omega_d, \quad \text{whenever } \varepsilon^2 \le \omega_d/\nu_d. \tag{6.48}$$

Note that (6.48) is a form of first-order Taylor expansion around u_d.

The following refines Theorem 10.

Proposition 2. *In the setting considered here,*

$$\mathbb{P}(1-\varepsilon_d \le U/u_d \le 1+\varepsilon_d) \to 1, \quad \textit{whenever } \varepsilon_d \gg 1/\sqrt{\nu_d}. \tag{6.49}$$

Proof. By mononicity in $\varepsilon_d > 0$, it is enough to show that when $\varepsilon_d^2 \le \omega_d/\nu_d$. Then to prove (6.49), as before, it suffices to show that

$$\text{(left tail) } \frac{\mathbb{P}(U_d \le (1-\varepsilon_d)u_d)}{\mathbb{P}(U_d \ge (1-\varepsilon_d)u_d)} \to 0 \quad \text{and} \quad \text{(right tail) } \frac{\mathbb{P}(U_d \ge (1+\varepsilon_d)u_d)}{\mathbb{P}(U_d \le (1+\varepsilon_d)u_d)} \to 0. \tag{6.50}$$

Left tail. As before, we can show that $b := L((1-\varepsilon_d)u_d)$ satisfies $b - d \to -\infty$, so that we may assume that $b < d-1$. Then using the fact that $u^\ell \psi(u) \le u_0^\ell \psi(u_0)$ for any $u \le u_0 \le u_\ell$, where $u_\ell = L^{-1}(\ell)$, we have

$$\frac{1}{c_d}\,\mathbb{P}(U_d \le (1-\varepsilon_d)u_d) = \int_0^{(1-\varepsilon_d)u_d} u^d \psi(u)\mathrm{d}u \tag{6.51}$$

$$\le ((1-\varepsilon_d)u_d)^b \psi((1-\varepsilon_d)) \int_0^{(1-\varepsilon_d)u_d} u^{d-b}\mathrm{d}u \tag{6.52}$$

$$= \frac{((1-\varepsilon_d)u_d)^{d+1}}{d-b+1}\psi((1-\varepsilon_d)u_d). \tag{6.53}$$

Taking the ratio of (6.53) to (6.21) (but replacing ε by ε_d), we obtain

$$\frac{\mathbb{P}(U_d \le (1-\varepsilon_d)u_d)}{\mathbb{P}(U_d \ge (1-\varepsilon_d)u_d)} \le \frac{1}{d-b+1}\frac{1-\varepsilon_d}{\varepsilon_d/2}\left(\frac{1-\varepsilon_d}{1-\varepsilon_d/2}\right)^d \frac{\psi((1-\varepsilon_d)u_d)}{\psi((1-\varepsilon_d/2)u_d)}. \tag{6.54}$$

As in (6.23) and (6.27), we apply a logarithm, and obtain the upper bound

$$\log \frac{1-\varepsilon_d}{\frac{\varepsilon_d}{2}(d-b+1)} - [d - L((1-\varepsilon_d/2)u_d)] \log \frac{1-\varepsilon_d/2}{1-\varepsilon_d}. \tag{6.55}$$

By (6.48), which is applicable by our assumption $\varepsilon_d^2 \le \omega_d/\nu_d$, we have

$$d - b = L(u_d) - L((1-\varepsilon_d)u_d) = \varepsilon_d \nu_d \pm |\varepsilon_d|\nu_d/\omega_d \sim \varepsilon_d \nu_d, \quad d \to \infty. \tag{6.56}$$

Similarly,

$$d - L((1-\varepsilon_d/2)u_d) \sim \tfrac{1}{2}\varepsilon_d \nu_d, \quad d \to \infty. \tag{6.57}$$

Also, note that $\varepsilon_d^2 \nu_d \to \infty$ by assumption. Hence, the first term in (6.55) is $\sim -\log(\varepsilon_d^2 \nu_d)$ while the second term (including sign) is $\sim -\frac{1}{4}\varepsilon_d^2 \nu_d$, so that the sum tends to $-\infty$.

Right tail. The treatment of the right tail is analogous, starting with (6.40) instead of (6.23). Details are omitted. □

In the following we examine the behavior of the normalizing constant c_d as $d \to \infty$.

Proposition 3. *In the setting considered here,*

$$\frac{1}{c_d} \sim \sqrt{\frac{2\pi}{\nu_d}} u_d^{d+1} \psi(u_d), \quad d \to \infty. \tag{6.58}$$

Proof. Let ε_d be such that $1/\nu_d \ll \varepsilon_d^2 \ll \omega_d/\nu_d$. Applying Proposition 2 and then performing a change of variables, we get

$$\frac{1}{c_d} = \int_0^\infty u^d \psi(u) \mathrm{d}u \sim \int_{(1-\varepsilon_d)u_d}^{(1+\varepsilon_d)u_d} u^d \psi(u) \mathrm{d}u \tag{6.59}$$

$$\sim \int_{-\varepsilon_d}^{\varepsilon_d} [(1+t)u_d]^d \psi[(1+t)u_d] u_d \mathrm{d}t \tag{6.60}$$

$$= u_d^{d+1} \psi(u_d) \int_{-\varepsilon_d}^{\varepsilon_d} (1+t)^d \frac{\psi[(1+t)u_d]}{\psi(u_d)} \mathrm{d}t. \tag{6.61}$$

As before,

$$\log \Big\{ (1+t)^d \frac{\psi[(1+t)u_d]}{\psi(u_d)} \Big\} = d \log(1+t) - \Lambda[(1+t)u_d] + \Lambda(u_d) \tag{6.62}$$

$$= d \int_0^t \frac{s}{1+s} \mathrm{d}s - \int_0^t u_d \Lambda'[(1+s)u_d] \mathrm{d}s \tag{6.63}$$

$$= \int_0^t \frac{1}{1+s} \Big\{ d - L[(1+s)u_d] \Big\} \mathrm{d}s. \tag{6.64}$$

Noting that $|s| \leq \varepsilon_d$, and using (6.48), we get

$$-s\nu_d - s\nu_d/\omega_d \leq d - L[(1+s)u_d] = L(u_d) - L[(1+s)u_d] \leq -s\nu_d + s\nu_d/\omega_d. \quad (6.65)$$

Hence,

$$\begin{aligned} -\nu_d(1+1/\omega_d)\int_0^t \frac{s}{1+s}\mathrm{d}s &\leq \int_0^t \frac{1}{1+s}\Big\{d - L[(1+s)u_d]\Big\}\mathrm{d}s \\ &\leq -\nu_d(1-1/\omega_d)\int_0^t \frac{s}{1+s}\mathrm{d}s, \end{aligned} \quad (6.66)$$

with

$$\int_0^t \frac{s}{1+s}\mathrm{d}s = \tfrac{1}{2}t^2 + O(t^3) = \tfrac{1}{2}t^2 + O(\varepsilon_d^3), \quad (6.67)$$

since $|t| \leq \varepsilon_d$, so that

$$\int_0^t \frac{1}{1+s}\Big\{d - L[(1+s)u_d]\Big\}\mathrm{d}s = -\tfrac{1}{2}t^2\nu_d + O(\varepsilon_d + 1/\omega_d)\varepsilon_d^2\nu_d, \quad (6.68)$$

where the big-O is uniform in $t \in [-\varepsilon_d, \varepsilon_d]$. We already took ε_d such that $(1/\omega_d)\varepsilon_d^2\nu_d \to 0$, and it is compatible to choose ε_d such that, in addition, $\varepsilon_d^3\nu_d \to 0$. When we do so, the remainder term above is $o(1)$, and in particular,

$$\int_0^t \frac{1}{1+s}\Big\{d - L[(1+s)u_d]\Big\}\mathrm{d}s = -\tfrac{1}{2}t^2\nu_d + o(1), \quad (6.69)$$

where the $o(1)$ term is uniform in $t \in [-\varepsilon_d, \varepsilon_d]$. With such a choice of ε_d, we continue our derivations above

$$\int_{-\varepsilon_d}^{\varepsilon_d}(1+t)^d\frac{\psi[(1+t)u_d]}{\psi(u_d)}\mathrm{d}t = \int_{-\varepsilon_d}^{\varepsilon_d}\exp\Big\{-\tfrac{1}{2}t^2\nu_d + o(1)\Big\}\mathrm{d}t \quad (6.70)$$

$$\sim \frac{1}{\sqrt{\nu_d}}\int_{-\varepsilon_d\sqrt{\nu_d}}^{\varepsilon_d\sqrt{\nu_d}}\exp\Big\{-\tfrac{1}{2}s^2\Big\}\mathrm{d}s \sim \frac{1}{\sqrt{\nu_d}}\sqrt{2\pi}, \quad (6.71)$$

since $\varepsilon_d\sqrt{\nu_d} \to \infty$ by assumption. □

We are finally equipped to establish a convergence in distribution for U_d.

Theorem 11. *In the setting considered here, $\sqrt{\nu_d}(U_d/u_d - 1)$ converges weakly to the standard normal distribution as $d \to \infty$.*

Example 1 (Continued). It can be checked that the same example of shape function ψ satisfies the conditions assumed here, with $uL'(u) \sim c\beta^2(\log u)^\alpha u^\beta$ as $u \to \infty$, so that

$$\nu_d = u_d L'(u_d) \sim \beta d, \quad d \to \infty. \tag{6.72}$$

Proof. Fix $r \in \mathbb{R}$ and let ε_d be as before. As in (6.61), we get

$$\mathbb{P}(\sqrt{\nu_d}(U_d/u_d - 1) \le r) = \mathbb{P}(U_d \le (1 + r/\sqrt{\nu_d})u_d) \tag{6.73}$$

$$\sim \mathbb{P}((1 - \varepsilon_d)u_d \le U_d \le (1 + r/\sqrt{\nu_d})u_d) \tag{6.74}$$

$$= c_d u_d^{d+1} \psi(u_d) \int_{-\varepsilon_d}^{r/\sqrt{\nu_d}} (1+t)^d \frac{\psi[(1+t)u_d]}{\psi(u_d)} \mathrm{d}t. \tag{6.75}$$

Again, as before,

$$\int_{-\varepsilon_d}^{r/\sqrt{\nu_d}} (1+t)^d \frac{\psi[(1+t)u_d]}{\psi(u_d)} \mathrm{d}t \sim \int_{-\varepsilon_d}^{r/\sqrt{\nu_d}} \exp\Big\{ -\tfrac{1}{2} t^2 \nu_d \Big\} \mathrm{d}t \tag{6.76}$$

$$= \frac{1}{\sqrt{\nu_d}} \int_{-\varepsilon_d\sqrt{\nu_d}}^{r} \exp\Big\{ -\tfrac{1}{2} s^2 \Big\} \mathrm{d}s \sim \frac{1}{\sqrt{\nu_d}} \sqrt{2\pi} \Phi(r), \tag{6.77}$$

where Φ is the standard normal distribution function. We then combine this with Proposition 3. □

6.3 Consequences for statistical modeling

While there is relatively little related work, a detailed comparison with (Sherlock and Elton, 2012) is in order. Sherlock and Elton focus on the non-compact case — corresponding to Section 6.2 here. They derive the same result as our Theorem 10 under different conditions. They require that $\eta(u) := \Lambda(\exp(u))$ is twice differentiable with $\eta''(u) \to \infty$ as $u \to \infty$, while our condition is a bit weaker than requiring that η is once differentiable with $\eta'(u)/u \to \infty$ and increasing. Note that their condition is equivalent to requiring that L is differentiable with $uL'(u) \to \infty$, a condition that arises in Section 6.2.2. Sherlock and Elton do not establish weak convergence, however, but they obtain other results. In particular, they establish concentration for a marginal of X_d and also for the maximum of the marginals.

In addition, they extend their results to the case of elliptical distributions under conditions on the eigenvalues of the scaling matrix.

What are possible consequences for statistical modeling? Because of the weak convergence of the sort established here, the behavior of U_d is asymptotically characterized solely by a few parameters of the underlying distribution. For example, if the conditions of Theorem 9 are fulfilled, then the distribution of U_d in the large-dimension limit ($d \to \infty$) only depends on u_* (irrelevant in practice because scale is typically estimated) and the behavior of ψ near u_*. In particular, whether $\psi(u) = \mathbb{I}\{u \leq 1\}$ or $\psi(u) = (2-u)\mathbb{I}\{u \leq 1\}$, in both cases, $d(1-U_d)$ converges weakly to the exponential distribution with rate 1. This means that, in order to even distinguish two such distributions with nontrivial accuracy, we require a sample of size that increases with d. (We did not attempt to quantify this further, although this is possible by framing the problem as a hypothesis testing problem.) A similar phenomenon arises with certain distributions with non-compact support, based on our Theorem 11. Thus, if the sample size is small relative to the dimension, very different shape functions (meaning, different ψ's) could yield indistinguishable models.

The flip side of this is a form of universality of the Gaussian distribution, in particular, in context such as Linear Discriminant Analysis (classification) or Gaussian Mixture Modeling (clustering). Surely, both LDA and GMM have computational advantages over other methods (the latter using the EM algorithm, for example). Beyond this important computational aspect, our results indicate that when the sample is small relative to the dimension, fitting a Gaussian model may be, in fact, indistinguishable from fitting another model base on a shape function having similar characteristics as the standard normal distribution that dictate the asymptotic behavior of U_d.

Acknowledgement

This chapter, in full, has been organized into the following paper: *Concentration of Measure for Radial Distributions and Consequences for Statistical Modeling*

(Ery Arias-Castro and Xiao Pu), and has been submitted for publication. The dissertation author is the corresponding author of this material.

Bibliography

Aggarwal, C. C., J. L. Wolf, P. S. Yu, C. Procopiuc, and J. S. Park (1999). Fast algorithms for projected clustering. In *ACM SIGMoD Record*, Volume 28, pp. 61–72. ACM.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control 19*(6), 716–723.

Anderson, T. (2004). An introduction to multivariate statistical analysis.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis* (Third ed.). Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ.

Anevski, D. (1994). *Estimating the derivative of a convex density.*

Arias-Castro, E. and X. Pu (2016). Concentration of measure for radial distributions and consequences for statistical modeling. *arXiv preprint arXiv:1607.07549*.

Arias-Castro, E. and X. Pu (2017). A simple approach to sparse clustering. *Computational Statistics & Data Analysis 105*, 217–228.

Arthur, D. and S. Vassilvitskii (2007). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1027–1035. Society for Industrial and Applied Mathematics.

Azizyan, M., A. Singh, and L. Wasserman (2013). Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. *Neural Information Processing Systems (NIPS)*.

Bagnoli, M. and T. C. Bergstrom (1989). Log-concave probability and its applications.

Balabdaoui, F. (2004). *Nonparametric estimation of a k-monotone density: A new asymptotic distribution theory.* Ph. D. thesis, University of Washington.

Balabdaoui, F. and C. R. Doss (2014). Inference for a mixture of symmetric

distributions under log-concavity. *arXiv preprint arXiv:1411.4708*.

Balabdaoui, F., K. Rufibach, and J. A. Wellner (2009). Limit distribution theory for maximum likelihood estimation of a log-concave density. *Annals of statistics 37*(3), 1299.

Benaglia, T., D. Chauveau, and D. R. Hunter (2009). An em-like algorithm for semi-and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics 18*(2), 505–526.

Bhattacharyya, S. and P. J. Bickel (2015). Adaptive estimation in elliptical distributions with extensions to high dimensions. Technical report, University of California, Berkeley.

Bickel, P. J. and J. Fan (1996). Some problems on the estimation of unimodal densities. *Statistica Sinica*, 23–45.

Bickel, P. J., C. A. Klaassen, Y. Ritov, and J. A. Wellner (1998). Efficient and adaptive estimation for semiparametric models.

Bickel, P. J. and E. Levina (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics*, 199–227.

Bordes, L., D. Chauveau, and P. Vandekerkhove (2007). A stochastic em algorithm for a semiparametric mixture model. *Computational Statistics & Data Analysis 51*(11), 5429–5443.

Bordes, L., S. Mottelet, and P. Vandekerkhove (2006). Semiparametric estimation of a two-component mixture model. *The Annals of Statistics 34*(3), 1204–1232.

Boucheron, S., G. Lugosi, and P. Massart (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Butucea, C. and P. Vandekerkhove (2014). Semiparametric mixtures of symmetric distributions. *Scandinavian Journal of Statistics 41*(1), 227–239.

Cai, T. T., Z. Ma, and Y. Wu (2013). Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics 41*(6), 3074–3110.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 2313–2351.

Carson, T. and R. Impagliazzo (2001). Hill-climbing finds random planted bisections. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pp. 903–909. Society for Industrial and Applied Mathematics.

Chan, Y.-b. and P. Hall (2010). Using evidence of mixed populations to select

variables for clustering very high-dimensional data. *Journal of the American Statistical Association 105*(490), 798–809.

Chang, G. T. and G. Walther (2007). Clustering with mixtures of log-concave distributions. *Computational Statistics & Data Analysis 51*(12), 6242–6251.

Chauveau, D. and V. T. L. Hoang (2016). Nonparametric mixture models with conditionally independent multivariate component densities. *Computational Statistics & Data Analysis 103*, 1–16.

Chauveau, D., D. R. Hunter, and M. Levine (2015). Semi-parametric estimation for conditional independence multivariate finite mixture models. *Statistics Surveys 9*, 1–31.

Cule, M. and R. Samworth (2010). Theoretical properties of the log-concave maximum likelihood estimator of a multidimensional density. *Electronic Journal of Statistics 4*, 254–270.

Cule, M., R. Samworth, and M. Stewart (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72*(5), 545–607.

d'Aspremont, A., L. El Ghaoui, M. I. Jordan, and G. R. Lanckriet (2007). A direct formulation for sparse pca using semidefinite programming. *SIAM review 49*(3), 434–448.

Davis, C. and W. M. Kahan (1970). The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis 7*(1), 1–46.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1–38.

Dettling, M. (2004). Bagboosting for tumor classification with gene expression data. *Bioinformatics 20*(18), 3583–3593.

Doss, C. R. and J. A. Wellner (2016a). Global rates of convergence of the mles of log-concave and s-concave densities. *The Annals of Statistics 44*(3), 954–981.

Doss, C. R. and J. A. Wellner (2016b). Mode-constrained estimation of a log-concave density. *arXiv preprint arXiv:1611.10335*.

Dümbgen, L. and K. Rufibach (2009). Maximum likelihood estimation of a log-concave density and its distribution function: Basic properties and uniform consistency. *Bernoulli 15*(1), 40–68.

Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper.

ACM SIGKDD explorations newsletter 4(1), 65–75.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Fraley, C. and A. E. Raftery (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association 97*(458), 611–631.

Fraley, C. and A. E. Raftery (2006). Mclust version 3: an r package for normal mixture modeling and model-based clustering. Technical report, DTIC Document.

Friedman, J. H. and J. J. Meulman (2004). Clustering objects on subsets of attributes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(4), 815–849.

Ghosh, D. and A. M. Chinnaiyan (2002). Mixture modelling of gene expression data from microarray experiments. *Bioinformatics 18*(2), 275–286.

Grenander, U. (1956). On the theory of mortality measurement: part ii. *Scandinavian Actuarial Journal 1956*(2), 125–153.

Groeneboom, P., G. Jongbloed, and J. A. Wellner (2001). Estimation of a convex function: characterizations and asymptotic theory. *Annals of Statistics*, 1653–1698.

Guo, J., E. Levina, G. Michailidis, and J. Zhu (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics 66*(3), 793–804.

Hall, P. and X.-H. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Annals of Statistics*, 201–224.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning*. New York: Springer.

Hettmansperger, T. and H. Thomas (2000). Almost nonparametric inference for repeated measures in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 62*(4), 811–825.

Horn, R. A. and C. R. Johnson (2012). *Matrix analysis*. Cambridge university press.

Hu, H., Y. Wu, and W. Yao (2016). Maximum likelihood estimation of the mixture of log-concave densities. *Computational Statistics & Data Analysis 101*, 137–147.

Hu, H., W. Yao, and Y. Wu (2017). The robust em-type algorithms for log-concave mixtures of regression models. *Computational Statistics & Data Analysis 111*, 14–26.

Hunter, D. R., S. Wang, and T. P. Hettmansperger (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics*, 224–251.

Jin, J., Z. T. Ke, and W. Wang (2015). Phase transitions for high dimensional clustering and related problems. *arXiv preprint arXiv:1502.06952*.

Jin, J. and W. Wang (2014). Important feature pca for high dimensional clustering. *arXiv preprint arXiv:1407.5241*.

Johnson, W. B. and J. Lindenstrauss (2001). *Handbook of the geometry of Banach spaces*, Volume 1. Elsevier.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, 295–327.

Johnstone, I. M. and A. Y. Lu (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association 104*(486).

Jolliffe, I. T., N. T. Trendafilov, and M. Uddin (2003). A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics 12*(3), 531–547.

Jongbloed, G. (1995). Three statistical inverse problems: Estimators-algorithms-asymptotics.

Kernighan, B. W. and S. Lin (1970). An efficient heuristic procedure for partitioning graphs. *Bell system technical journal 49*(2), 291–307.

Khas'minskii, R. (1979). A lower bound on the risks of non-parametric estimates of densities in the uniform metric. *Theory of Probability & Its Applications 23*(4), 794–798.

Kou, J. (2014). Estimating the number of clusters via the gud statistic. *Journal of Computational and Graphical Statistics 23*(2), 403–417.

Krauthgamer, R., B. Nadler, and D. Vilenchik (2015). Do semidefinite relaxations solve sparse pca up to the information limit? *The Annals of Statistics 43*(3), 1300–1322.

Lam, C. and J. Fan (2009). Sparsistency and rates of convergence in large covariance matrix estimation. *Annals of statistics 37*(6B), 4254.

Ledoux, M. (2005). *The concentration of measure phenomenon*. Number 89. American Mathematical Soc.

Lee, A. B., D. Luca, and K. Roeder (2010). A spectral graph approach to discovering genetic ancestry. *The annals of applied statistics 4*(1), 179.

Levina, E., A. Rothman, and J. Zhu (2008). Sparse estimation of large covariance matrices via a nested lasso penalty. *The Annals of Applied Statistics 2*(1), 245–263.

Levine, M., D. Hunter, and D. Chauveau (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika 98*(2), 403–416.

Lewis, D. D. (1998). Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pp. 4–15. Springer.

Liu, J. S., J. L. Zhang, M. J. Palumbo, and C. E. Lawrence (2003). Bayesian clustering with variable and transformation selections. *Bayesian Statistics 7*, 245–275.

Lounici, K. (2013). Sparse principal component analysis with missing observations. In *High dimensional probability VI*, pp. 327–356. Springer.

Ma, Z. (2013). Sparse principal component analysis and iterative thresholding. *The Annals of Statistics 41*(2), 772–801.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Volume 1, pp. 281–297. Oakland, CA, USA.

Mallapragada, P. K., R. Jin, and A. Jain (2010). Non-parametric mixture models for clustering. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, pp. 334–343. Springer.

Mallows, C. L. (1973). Some comments on c p. *Technometrics 15*(4), 661–675.

McLachlan, G. and T. Krishnan (2007). *The EM algorithm and extensions*, Volume 382. John Wiley & Sons.

McLachlan, G. and D. Peel (2000). Mixtures of factor analyzers. *Finite Mixture Models*, 238–256.

McNeil, D. R. (1977). Interactive data analysis.

Meyer, M. C. and M. Woodroofe (2004). Consistent maximum likelihood esti-

mation of a unimodal density using shape restrictions. *Canadian Journal of Statistics 32*(1), 85–100.

Muirhead, R. J. (2009). *Aspects of multivariate statistical theory*, Volume 197. John Wiley & Sons.

Pan, W. and X. Shen (2007). Penalized model-based clustering with application to variable selection. *The Journal of Machine Learning Research 8*, 1145–1164.

Paul, D. (2007). Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Statistica Sinica 17*(4), 1617.

Pu, X. and E. Arias-Castro (2017). Semiparametric estimation of symmetric mixture models with monotone and log-concave densities. *arXiv preprint arXiv:1702.08897*.

Raftery, A. E. and N. Dean (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association 101*(473), 168–178.

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

Rockafellar, R. T. (2015). *Convex analysis*. Princeton university press.

Rufibach, K. (2006). *Log-concave density estimation and bump hunting for IID observations*. Ph. D. thesis.

Sherlock, C. and D. Elton (2012). A class of spherical and elliptical distributions with gaussian-like limit properties. *Journal of Probability and Statistics 2012*.

Street, W. N., W. H. Wolberg, and O. L. Mangasarian (1993). Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, pp. 861–870. International Society for Optics and Photonics.

Sun, W., J. Wang, and Y. Fang (2012). Regularized k-means clustering of high-dimensional data and its asymptotic consistency. *Electronic Journal of Statistics 6*, 148–167.

Tamayo, P., D. Scanfeld, B. L. Ebert, M. A. Gillette, C. W. Roberts, and J. P. Mesirov (2007). Metagene projection for cross-platform, cross-species characterization of global transcriptional states. *Proceedings of the National Academy of Sciences 104*(14), 5959–5964.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288.

Tibshirani, R., G. Walther, and T. Hastie (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63*(2), 411–423.

Verzelen, N. and E. Arias-Castro (2014). Detection and feature selection in sparse mixture models. *arXiv preprint arXiv:1405.1478*.

Vu, V. Q. and J. Lei (2012). Minimax rates of estimation for sparse pca in high dimensions. *arXiv preprint arXiv:1202.0786*.

Vu, V. Q. and J. Lei (2013). Minimax sparse principal subspace estimation in high dimensions. *The Annals of Statistics 41*(6), 2905–2947.

Wang, S. and J. Zhu (2008). Variable selection for model-based high-dimensional clustering and its application to microarray data. *Biometrics 64*(2), 440–448.

Wegman, E. J. (1970). Maximum likelihood estimation of a unimodal density function. *The Annals of Mathematical Statistics 41*(2), 457–471.

Witten, D. M. and R. Tibshirani (2010). A framework for feature selection in clustering. *Journal of the American Statistical Association 105*(490).

Witten, D. M., R. Tibshirani, and T. Hastie (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, kxp008.

Woodroofe, M. and J. Sun (1993). A penalized maximum likelihood estimate of f (0+) when f is non-increasing. *Statistica Sinica*, 501–515.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, 95–103.

Xie, B., W. Pan, and X. Shen (2008). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic journal of statistics 2*, 168.

Yousefi, M. R., J. Hua, C. Sima, and E. R. Dougherty (2010). Reporting bias when using real data sets to analyze classification performance. *Bioinformatics 26*(1), 68–76.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of computational and graphical statistics 15*(2), 265–286.