Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Robust Spectral Clustering Algorithm for Sub-Gaussian Mixture Models with Outliers

Prateek R. Srivastava, Purnamrita Sarkar, Grani A. Hanasusanto

To cite this article:
Prateek R. Srivastava, Purnamrita Sarkar, Grani A. Hanasusanto (2023) A Robust Spectral Clustering Algorithm for Sub-Gaussian Mixture Models with Outliers. Operations Research 71(1):224-244. https://doi.org/10.1287/opre.2022.2317

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
Vol. 71, No. 1, January–February 2023, pp. 224–244
ISSN 0030-364X (print), ISSN 1526-5463 (online)

Crosscutting Areas

A Robust Spectral Clustering Algorithm for Sub-Gaussian Mixture Models with Outliers

Prateek R. Srivastava,[a,*] Purnamrita Sarkar,[b] Grani A. Hanasusanto[c]

[a] Graduate Program in Operations Research and Industrial Engineering (ORIE), University of Texas at Austin, Austin, Texas 78712; [b] Department of Statistics and Data Sciences, University of Texas at Austin, Austin, Texas 78712; [c] Graduate Program in Operations Research and Industrial Engineering, University of Texas at Austin, Austin, Texas 78712
*Corresponding author

Contact: prateekrs@utexas.edu, https://orcid.org/0000-0002-1188-4206 (PRS); purna.sarkar@austin.utexas.edu, https://orcid.org/0000-0002-6542-1215 (PS); grani.hanasusanto@utexas.edu, https://orcid.org/0000-0003-4900-2958 (GAH)

Received: January 27, 2020
Revised: January 31, 2021; November 28, 2021
Accepted: May 2, 2022
Published Online in Articles in Advance: July 11, 2022

Area of Review: Machine Learning and Data Science.

https://doi.org/10.1287/opre.2022.2317

Abstract. We consider the problem of clustering data sets in the presence of arbitrary outliers. Traditional clustering algorithms such as k-means and spectral clustering are known to perform poorly for data sets contaminated with even a small number of outliers. In this paper, we develop a provably robust spectral clustering algorithm that applies a simple rounding scheme to denoise a Gaussian kernel matrix built from the data points and uses vanilla spectral clustering to recover the cluster labels of data points. We analyze the performance of our algorithm under the assumption that the "good" data points are generated from a mixture of sub-Gaussians (we term these "inliers"), whereas the outlier points can come from any arbitrary probability distribution. For this general class of models, we show that the misclassification error decays at an exponential rate in the signal-to-noise ratio, provided the number of outliers is a small fraction of the inlier points. Surprisingly, this derived error bound matches with the best-known bound for semidefinite programs (SDPs) under the same setting without outliers. We conduct extensive experiments on a variety of simulated and real-world data sets to demonstrate that our algorithm is less sensitive to outliers compared with other state-of-the-art algorithms proposed in the literature.

Funding: G. A. Hanasusanto was supported by the National Science Foundation Grants NSF ECCS-1752125 and NSF CCF-2153606. P. Sarkar gratefully acknowledges support from the National Science Foundation Grants NSF DMS-1713082, NSF HDR-1934932 and NSF 2019844.
Supplemental Material: The online appendix is available at https://doi.org/10.1287/opre.2022.2317.

Keywords: spectral clustering • sub-Gaussian mixture models • kernel methods • semidefinite programming • outlier detection • asymptotic analysis

1. Introduction

Clustering is a fundamental problem in unsupervised learning, with application domains ranging from evolutionary biology, market research, and medical imaging to recommender systems and social network analysis, etc. In this paper, we consider the problem of clustering n independent and identically distributed inlier data points in d-dimensional space from a mixture of r sub-Gaussian probability distributions with unknown means and covariance matrices in the presence of arbitrary outlier data points. Given a sample data set consisting of these inlier and outlier points, the objective of our inference problem is to recover the latent cluster memberships for the set of inlier points and, additionally, to identify the outlier points in the data set.

Sub-Gaussian mixture models (SGMMs) are an important class of mixture models that provide a distribution-free approach for analyzing clustering algorithms and encompass a wide variety of fundamental clustering models, such as (i) spherical and general Gaussian mixture models (GMMs), (ii) stochastic ball models (Iguchi et al. 2015, Kushagra et al. 2017), which are mixture models whose components are isotropic distributions supported on unit ℓ_2-balls, and (iii) mixture models with component distributions that have a bounded support, as its special cases.

Taking the clustering objective and tractability of algorithms into consideration, several different solution schemes based on Lloyd's algorithm (Lloyd 1982), expectation maximization (Dempster et al. 1977), method of moments (Pearson 1936, Bickel et al. 2011), spectral methods (Dasgupta 1999, Vempala and Wang 2004), linear programming (Awasthi et al. 2015), and semidefinite programming (Peng and Wei 2007, Mixon et al. 2017, Yan and Sarkar 2021) have been proposed for clustering SGMMs. Among these different algorithms, Lloyd's algorithm, which is a popular heuristic

to solve the k-means clustering problem, is arguably the most widely used. When the data lie on a low dimensional manifold, a popular alternative is spectral clustering, which applies k-means on the top eigenvectors of a suitably normalized kernel similarity matrix (Shi and Malik 2000, Ng et al. 2002, Von Luxburg 2007, Von Luxburg et al. 2008, Schiebinger et al. 2015, Amini and Razaee 2021).

Despite their popularity, the performances of vanilla versions of both k-means clustering and spectral clustering are known to deteriorate in the presence of noise (Li et al. 2007, Bojchevski et al. 2017, Zhang and Rohe 2018). Figure 1 illustrates a simple example where the two algorithms fail in the presence of outlier points.

1.1. Our Contributions

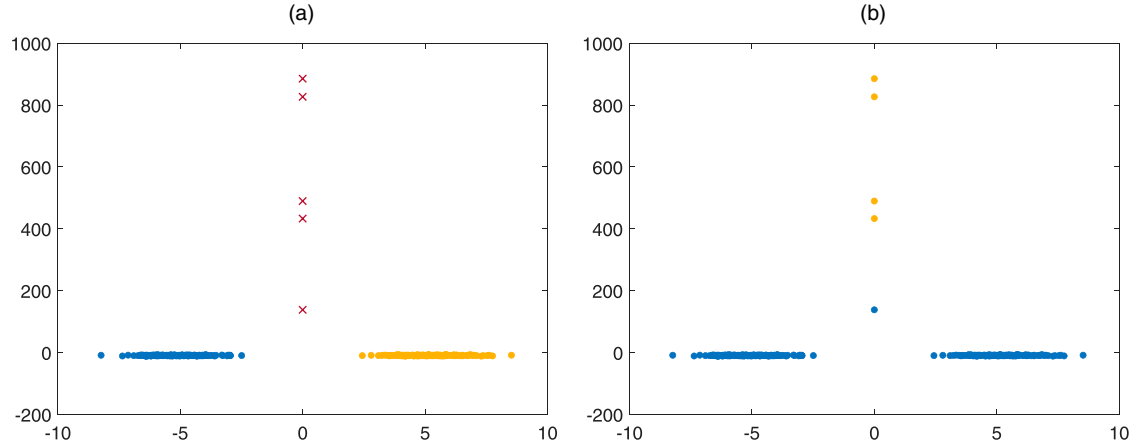
In this paper, we consider the joint kernel clustering and outlier detection problem under a SGMM setting assuming an arbitrary probability distribution for the set of outlier points. First, we formulate the exact kernel clustering problem with outliers and propose a robust SDP-based relaxation for the problem, which is applied after the data have been projected onto the top $r-1$ principal components (when $d > r$). This projection step not only helps tighten our theoretical bounds but also yields better empirical results when the dimensionality is large.

Because SDP formulations do not usually scale well to large problems, we propose a linear programming relaxation that essentially rounds the kernel matrix, on which we apply spectral clustering. In some sense, this algorithm is reminiscent of building a nearest neighbor graph from the data and applying spectral clustering on it. In the literature, k-nearest neighbor graphs have found applications in several machine-learning algorithms (Cover and Hart 1967, Altman 1992, Hastie and Tibshirani 1996, Ding and He 2004, Franti et al. 2006) and have been analyzed in the context of density-based clustering algorithms (Du et al. 2016, Verdinelli and Wasserman 2018) and subspace clustering (Heckel and Bölcskei 2015).

In general, kernel-based methods are harder to analyze compared with distance-based algorithms because they involve analyzing nonlinear feature transformations through the kernel function. In this work, we show that with high probability our algorithm recovers true cluster labels with small error rates for the set of inlier points, provided that there is a reasonable separation between the cluster centers and the number of outliers is not large. An interesting theoretical result that emerges from our analysis is that the error rate obtained for our spectral clustering algorithm decays exponentially in the square of the signal-to-noise ratio for the case when no outliers are present, which matches with the best-known theoretical error bound for SDP formulations (Fei and Chen 2018) under the SGMM setting.

Empirically, we observe a similar trend in the performances of robust spectral clustering and our proposed robust SDP-based formulation on real-world datasets, whereas the first is orders of magnitude faster. This is quite surprising, because in other model scenarios like the Stochastic Block Model (Holland et al. 1983), SDPs have been proven to return clusterings correlated to the ground truth in sparse data regimes (Guédon and Vershynin 2016, Montanari and

Figure 1. (Color online) k-Means++ and the Spectral Clustering Algorithm Proposed by Vempala and Wang (2004) are not Robust to the Outliers



Notes. The original data set consists of inlier data points (marked as solid circles) drawn from a mixture of two Gaussian distributions with means $\boldsymbol{\mu}_1 = [-5,0]^\top$, $\boldsymbol{\mu}_2 = [5,0]^\top$, covariance matrices $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \mathbf{I}_2$, and number of points $n_1 = n_2 = 150$. There are $m = 5$ outlier points generated on the y-axis, which are marked as red x's. In the clustering obtained from both the algorithms, the original clusters are merged into one, and the second cluster is comprised entirely of the outlier data points. (a) Original Data set; (b) clustering result obtained from k-means++ and spectral clustering (Vempala and Wang 2004).

Sen 2016), whereas only regularized variants of spectral clustering (Amini et al. 2013, Le et al. 2015, Joseph and Yu 2016, Zhang and Rohe 2018) work in these parameter regimes. However, to be fair, empirically we see that SDP is less sensitive to hyperparameter misspecification. We now summarize the main contributions of this paper.

1. We derive an exact formulation for the kernel clustering problem with outliers and obtain its SDP-based convex relaxation in the presence of outliers in the data set. Unlike previously proposed robust SDP formulations (Yan and Sarkar 2016, Rujeerapaiboon et al. 2019), our robust SDP formulation does not require prior knowledge of the number of clusters, the number of outliers, or cluster cardinalities.

2. We propose an efficient algorithm based on rounding and spectral clustering, which is provably robust. Specifically, we show that, provided the number of outliers is small compared with the inlier points, the error rate for our algorithm decays exponentially in the square of the signal-to-noise ratio. This error rate is consistent with the best-known theoretical error bound for SDP formulations (Fei and Chen 2018, Giraud and Verzelen 2018).

Although an extensive amount of work has been done previously to analyze spectral methods in the context of GMMs (Dasgupta 1999, Vempala and Wang 2004, Löffler et al. 2021), to the best of our knowledge, no prior theoretical work has been done to analyze robust spectral clustering algorithms for the nonparametric and more general SGMM setting (with or without outliers).

1.2. Related Work

Several previous works (Cuesta-Albertos et al. 1997, Li et al. 2007, Forero et al. 2012, Bojchevski et al. 2017, Zhang and Rohe 2018) have proposed robust variants of k-means and spectral clustering algorithms; however, they do not provide any recovery guarantees. Recently, there has been a focus on developing robust algorithms based on semidefinite programming and analyzing them for special cases of SGMMs. Kushagra et al. (2017) developed a robust reformulation of the k-means clustering SDP proposed by Peng and Wei (2007) and derived exact recovery guarantees under arbitrary (not necessarily isotropic) and stochastic ball model settings using a primal-dual certificate. On a related note, Rujeerapaiboon et al. (2019) also obtained a robust SDP-based clustering solution by minimizing the k-means objective subject to explicit cardinality constraints on the clusters as well as the set of outlier points. Besides the SGMM setting, robust clustering algorithms have been proposed for the related problem of subspace clustering where similar theoretical guarantees have been obtained (Soltanolkotabi and Candés 2012, Wang and Xu 2013, Soltanolkotabi et al. 2014, Heckel and Bölcskei 2015, Heckel et al. 2017, Wang et al. 2018) as well as for some other model settings (Vinayak and Hassibi 2016, Yan and Sarkar 2016). Particularly relevant to us is the work of Yan and Sarkar (2016), who compared the robustness of kernel clustering algorithms based on SDPs and spectral methods. However, they analyzed the algorithms for the mixture model introduced by El Karoui (2010), which assumed the data to be generated from a low-dimensional signal in a high-dimensional noise setting. Intuitively, in this setting, the signal-to-noise ratio, defined as the ratio of the minimum separation between cluster centers ($\Delta_{\min}$) to the largest spectral norm ($\sigma_{\max}$) of the covariance matrices of the mixture components, grows as $\sqrt{d}$. These authors showed that without outliers, the SDP-based algorithm is strongly consistent, that is, it achieves exact recovery, whereas kernel SVD algorithm is weakly consistent, that is, the fraction of misclassified data points go to zero in the limit as long as d increases polynomially in N, the total number of points. Note that in typical mixture models, the number of dimensions, although arbitrarily large, stay fixed, and there is a possibly small yet nonvanishing Bayes error rate, which is more realistic.

For the no outliers setting, an extensive amount of work has been done to obtain theoretical guarantees on the performances of various clustering algorithms under different distributional assumptions about the underlying data generation process. For the Gaussian mixture model setting, Dasgupta (1999) was among the first to obtain theoretical guarantees for a random projections-based clustering algorithm that is able to learn the parameters of mixture model provided the minimum separation between cluster centers $\Delta_{\min} = \Omega(\sqrt{d}\sigma_{\max})$. Using distance concentration arguments based on the isoperimetric inequality, Arora and Kannan (2001) improved the minimum separation to $\Delta_{\min} = \Omega(d^{1/4}\sigma_{\max})$. For the special case of a mixture r spherical Gaussians, Vempala and Wang (2004) showed that for their spectral algorithm the separation can be further reduced to $\Delta_{\min} = \Omega((r\log d)^{1/4}\sigma_{\max})$, which, ignoring the logarithmic factor in d, is essentially independent of the dimension of the problem. These results are generalized and extended further in subsequent works of Kumar and Kannan (2010) and Awasthi and Sheffet (2012). For a distribution-free model described in terms of the proximity conditions considered in Kumar and Kannan (2010), Li et al. (2020) obtained guarantees for the Peng and Wei (2007) k-means SDP relaxation. Under the stochastic ball model setting, Awasthi et al. (2015) obtained exact recovery guarantees for linear programming and SDP-based formulations for k-median and k-means clustering problems using a

primal-dual certificate argument. Extending the results of Awasthi et al. (2015), Mixon et al. (2017) showed that for a mixture of sub-Gaussians, the SDP-based formulation proposed in Peng and Wei (2007) guarantees good approximations to the true cluster centers provided the minimum distance between cluster centers $\Delta_{\min} = \Omega(r\sigma_{\max})$. Under a similar separation condition, Yan and Sarkar (2021) also obtained recovery guarantees for a kernel-based SDP formulation under the SGMM setting. Most pertinent to us is the recent result obtained by Fei and Chen (2018), who showed that for a minimum separation of $\Delta_{\min} = \Omega(\sqrt{r}\sigma_{\max})$ the misclassification error rate of a SGMM with equal-sized clusters decays exponentially in the square of the signal-to-noise ratio. Another analogous result for the SDP formulation proposed by Peng and Wei (2007) has been obtained by Giraud and Verzelen (2018). Very recently, we also became aware of the result obtained by Löffler et al. (2021), who obtained an exponentially decaying error rate for a spectral clustering algorithm for the special case of spherical Gaussians with identity covariance matrices. However, in order for their result to hold with high probability, they required the minimum separation between cluster centers to go to infinity. In addition, their proposed algorithm can easily be shown to fail in the presence of outliers, as discussed in greater detail in Section 4. For a clear comparison of our work with these notable works, we have included Table 1.

In addition to the clustering literature where data are typically drawn i.i.d. from a mixture distribution, spectral and SDP relaxations for hard combinatorial optimization problems have also received significant attention in graph partitioning and community detection literature (Goemans and Williamson 1995, McSherry 2001, Newman 2006, Rohe et al. 2011, Sussman et al. 2012, Fishkind et al. 2013, Qin and Rohe 2013, Guédon and Vershynin 2016, Amini and Levina 2018, Yan et al. 2018).

1.3. Paper Organization

The remainder of the paper is structured as follows. In Section 2, we introduce the notation used in the paper and describe the problem setup for sub-Gaussian mixture models with outliers. In Section 3, we obtain the formulation for the kernel clustering problem with outliers and derive its SDP and LP relaxations that recover denoised versions of the kernel matrix. In addition, we also discuss the details of the clustering algorithm that obtains cluster labels from this denoised matrix. Section 4 summarizes the main theoretical findings for our clustering algorithm, provides an overview of the proof techniques used, and contrasts our results with the existing results in the literature. Section 5 presents experimental results for several simulated and real-world data sets. Technical details of proofs for the main theorems are deferred to the online appendix.

2. Notation and Problem Setup

In this section, we introduce the notation used in this article and explain the formal setup of the kernel clustering problem for sub-Gaussian mixture models with outliers.

2.1. Notation

For any $n \in \mathbb{N}$, we define $[n]$ as the index set $\{1, \ldots, n\}$. We use uppercase boldfaced letters such as $\mathbf{A}, \mathbf{B}$ to denote matrices and lowercase boldfaced letters such

Table 1. Notable Related Works, Separation, Failure Probabilities, and Error Rates

Paper	SNR	Recovery type	Algorithm	Outliers	Failure probability	Error Rate
Vempala and Wang (2004)	$\Omega(r\log n)^{1/4}$	Exact	Spectral	No	o(1)	NA
Kumar and Kannan (2010)	$\Omega(r\cdot\text{polylog}(n))$	Exact	Spectral	No	o(1)	NA
Awasthi and Sheffet (2012)	$\Omega(\sqrt{r}\cdot\text{polylog}(n))$	Exact	Spectral	No	$o(1)$	NA
Lu and Zhou (2016)	$\Omega(r)$	Approx	Lloyd's algorithm initialized with Spectral Clustering	No	$e^{-\Omega(\text{SNR})}$	$e^{-\Omega(\text{SNR}^2)}$
	$\Omega\sqrt{\log(n)}$	Exact			$o(1)$	NA
Mixon et al. (2017)	$\Omega(r)$	Approx	SDP	No	o(1)	$\frac{1}{\text{SNR}^2}$
Fei and Chen (2018)	$\Omega(r)$	Approx	SDP	No	$o(1)$	$e^{-\Omega(\text{SNR}^2)}$
	$\Omega(r+\log(n))$	Exact				NA
Giraud and Verzelen (2018)	$\Omega(r^{1/2})$	Approx	SDP	No	o(1)	$e^{-\Omega(\text{SNR}^2)}$
Löffler et al. (2021)	$\Omega(r)$	Approx	Spectral	No	o(1)	$\frac{1}{\text{SNR}^2}$
	∞	Approx			$e^{-\Omega(\text{SNR})}$	$e^{-\Omega(\text{SNR}^2)}$
This paper	$\Omega(\sqrt{\min(d,r)})$	Approx	Spectral	Yes	o(1)	$e^{-\Omega(\text{SNR}^2)}$
	$\tilde{\Omega}\left(\sqrt{2+\eta}(\min(d,r))^{1/4}\right)$					$\frac{1}{\text{SNR}^\eta}$

Notes. For all methods that establish exact recovery, we have used NA as error rate. The $\tilde{\Omega}$ is used to hide a logarithmic factor in SNR.

as $\mathbf{u}, \mathbf{v}$ to denote vectors. For any matrix $\mathbf{A}$, $\mathrm{Tr}(\mathbf{A})$ denotes its trace, with A_{ij} its (i, j)-th entry, and $\mathrm{diag}(\mathbf{A})$ represents the column vector of its diagonal elements. We define $\mathrm{Diag}(\mathbf{v})$ to be a diagonal matrix with vector $\mathbf{v}$ on its main diagonal. We consider different matrix norms in our analysis. For a matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$, the operator norm $\|\mathbf{A}\|_2$ represents the largest singular value of $\mathbf{A}$, the Frobenius norm $\|\mathbf{A}\|_F = (\sum_{ij} A_{ij}^2)^{1/2}$ and ℓ_1-norm $\|\mathbf{A}\|_1 = \sum_{ij}|A_{ij}|$. For two matrices $\mathbf{A}, \mathbf{B}$ of same dimensions, the inner product between $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\langle \mathbf{A}, \mathbf{B}\rangle := \mathrm{Tr}(\mathbf{A}^\top \mathbf{B}) = \sum_{ij} A_{ij}B_{ij}$. We represent the n-dimensional vector of all ones by $\mathbf{1}_n$, the $n \times n$ matrix of all ones by $\mathbf{E}_n$, the $n \times n$ identity matrix by $\mathbf{I}_n$, and $n \times m$ matrix of all zeros by $\mathbf{0}_{n\times m}$. We define $\mathbf{e}_i$ to be the i-th standard basis vector whose i-th coordinate is 1, and all other coordinates are 0. We use $\mathbb{S}_n^+$ to denote the cone of $n \times n$ symmetric positive semidefinite matrices. Furthermore, we say that an $n \times n$ matrix $\mathbf{X} \succeq \mathbf{0}$ if and only if $\mathbf{X} \in \mathbb{S}_N^+$.

For the asymptotic analysis, we use standard notations like o, O, Ω and Θ to represent rates of convergence. We also use standard probabilistic order notations like O_p and o_p (see Van der Vaart 2000 for more details). We define $x \lesssim y$ to denote $x \le cy$, where c is some positive constant. We use $\tilde{O}$ to denote O with logarithmic dependence on the model parameters.

2.2. Problem Setup

We consider a generative model that generates a set of n independent and identically distributed inlier points, denoted by $\mathcal{I}$, from a mixture of r sub-Gaussian probability distributions (Vershynin 2012) $\{\mathcal{D}_k\}_{k=1}^r$. The set $\mathcal{O}$ of outlier points can come from arbitrary distributions with $|\mathcal{O}| = m$. Given the observed data matrix $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]^\top \in \mathbb{R}^{N\times d}$ consisting of these $N := n + m$ points in d-dimensional space, the task is to recover the latent cluster labels for the set of inlier points $\mathcal{I}$ and identify the outliers $\mathcal{O}$ in the data set.

For the set of inlier points, let $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_r)$, where $\boldsymbol{\pi} \ge \mathbf{0}$ and $\boldsymbol{\pi}^\top \mathbf{1}_r = 1$ denote the mixing weights associated with the r sub-Gaussian probability distributions in the mixture model such that $\pi_{\max} = \max_{k\in[r]} \pi_k$ and $\pi_{\min} = \min_{k\in[r]} \pi_k$. Assume that $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_r \in \mathbb{R}^d$ represent the means of r clusters from which the data points are generated. Under the SGMM model, for each point $i \in \mathcal{I}$, first a label $\phi_i \in \{1, \ldots, r\}$ is generated from a multinomial($\boldsymbol{\pi}$), where $\boldsymbol{\pi}$ is a r-dimensional vector denoting the cluster proportions. We define the true cluster membership matrix $\mathbf{Z}^0 \in \{0,1\}^{N\times r}$ such that $Z_{ik}^0 = 1$ if and only if point $i \in \mathcal{I}$ and $\phi_i = k$. Thus, assuming $Z_{ik}^0 = 1$, observation $\mathbf{y}_i$ is generated from distribution $\mathcal{D}_k$ with the following form:

$$\mathbf{y}_i := \boldsymbol{\mu}_k + \boldsymbol{\xi}_i,$$

where $\boldsymbol{\xi}_i$ is a mean zero sub-Gaussian random vector with σ_k^2 defined as the largest eigenvalue of its second moment matrix and $\sigma_{\max} := \max_{k\in[r]} \sigma_k$. We represent the k-th cluster by $\mathcal{C}_k := \{i \in \mathcal{I} : \phi_i = k\}$ and its cardinality by $n_k := |\mathcal{C}_k|$. The separation between any pair of clusters k and l is defined as $\Delta_{kl} := \|\boldsymbol{\mu}_k - \boldsymbol{\mu}_l\|_2$, with the minimum and maximum separation denoted respectively as $\Delta_{\min} := \min_{k\neq l} \Delta_{kl}$ and $\Delta_{\max} := \max_{k\neq l} \Delta_{kl}$. In our analysis, an important quantity of interest is the signal-to-noise ratio, which, based on Fei and Chen (2018), is defined as

$$\mathrm{SNR} := \frac{\Delta_{\min}}{\sigma_{\max}}. \tag{1}$$

Without loss of generality, we assume that the points in $\mathbf{Z}^0$ are ordered such that the inliers and outliers are indexed together. Within the set of inliers again, we further assume that the points belonging to the same cluster are indexed together. Thus, the true clustering matrix $\mathbf{X}^0 = \mathbf{Z}^0\mathbf{Z}^{0\top}$ is a block diagonal matrix with $X_{ij}^0 = 1$ if i and j belong to the same cluster and 0 otherwise. For our algorithm, we use the Gaussian kernel matrix $\mathbf{K} \in [0,1]^{N\times N}$, whose (i, j)-th entry $K_{ij} := \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|^2}{2\theta^2}\right)$ defines the similarity between points i and j for some scaling parameter θ.

3. Robust Kernel Clustering Formulation

Yu and Shi (2003) showed that the normalized k-cut problem is equivalent to the following trace maximization problem $\mathrm{Tr}(\mathbf{Z}^\top \mathbf{K} \mathbf{Z})$, where $\mathbf{Z}$ is a scaled cluster membership matrix. In their seminal paper, Dhillon et al. (2004) proved the equivalence between the kernel k-means and normalized k-cut problem. Based on Dhillon et al. (2004) and Yu and Shi (2003), Yan and Sarkar (2016) proposed a SDP relaxation for the kernel clustering problem under the assumption of equal-sized clusters. Yan and Sarkar (2021) further extended the kernel clustering formulation to unequal-sized clusters for analyzing the community detection problem in the presence node covariate information. Their formulation, which is derived from the SDP formulation for the k-means clustering problem (Peng and Wei 2007), however, does not account for possible outliers in the data set.

In this section, we first consider an exact formulation for the kernel clustering problem with equal-sized clusters and no outliers. We then extend this formulation to incorporate the case where cluster sizes may be unequal as well as unknown, and outliers are present in the data set. Finally, we use the idea of "lifting" and "relaxing" to obtain two efficient algorithms based on tractable SDP and spectral relaxations

for this exact formulation:

$$\begin{aligned} \underset{\mathbf{Z}}{\text{maximize}} \quad & \langle \mathbf{K}, \mathbf{Z}\mathbf{Z}^\top \rangle \\ \text{subject to} \quad & \mathbf{Z} \in \{0,1\}^{n\times r} \\ & \sum_{k\in[r]} Z_{ik} = 1 \quad \forall i = 1,\ldots,n \\ & \sum_{i\in[n]} Z_{ik} = \frac{n}{r} \quad \forall k = 1,\ldots,r \end{aligned} \tag{2}$$

The optimization formulation in (2) represents the kernel clustering problem without outliers that aims to maximize the sum of within-cluster similarities subject to assignment constraints that require each data point i to belong to exactly one cluster and cardinality constraints that assume all clusters to be equal-sized with exactly $\frac{n}{r}$ (assumed to be integral) data points in each cluster. For the case where the clusters are required to be equal-sized, the cardinality constraints in (2) can be equivalently expressed in an aggregated form by requiring $\langle \mathbf{E}_n, \mathbf{Z}\mathbf{Z}^\top \rangle = \frac{n^2}{r}$.

In general, however, the clusters are seldom equal-sized; in addition, their exact cardinalities are also seldom known in practice. However, if cardinality constraints are dropped from the formulation, the optimal solution $\mathbf{Z}^*$ assigns all points to a single cluster. A natural way to overcome this issue would be to maximize $\langle \mathbf{K} - \gamma \mathbf{E}_n, \mathbf{Z}\mathbf{Z}^\top \rangle$ for $\gamma \in (0, 1)$. Note that for a valid cluster membership matrix, $\mathbf{Z}$, $\langle \mathbf{E}_n, \mathbf{Z}\mathbf{Z}^\top \rangle = \frac{n^2}{r}$ represents its minimum value, which is achieved exactly when all of the clusters are equal-sized. Thus, the penalized objective function essentially tries to find clusters that are balanced.

We extend the formulation in (2) to account for possible outliers in the data set by relaxing the assignment constraint on each data point to belong to either exactly one cluster (if the data point is an inlier) or no cluster (if the data point is an outlier). The resulting exact formulation for the kernel clustering problem with outliers is a binary quadratic program and is shown in (3).

$$\begin{aligned} \underset{\mathbf{Z}}{\text{maximize}} \quad & \langle \mathbf{K} - \gamma \mathbf{E}_N, \mathbf{Z}\mathbf{Z}^\top \rangle \\ \text{subject to} \quad & \mathbf{Z} \in \{0,1\}^{N\times r} \\ & \mathbf{Z}\mathbf{1}_r \le \mathbf{1}_N. \end{aligned} \tag{3}$$

$$\begin{aligned} \underset{\mathbf{X}}{\text{maximize}} \quad & \langle \mathbf{K} - \gamma \mathbf{E}_N, \mathbf{X} \rangle \\ \text{subject to} \quad & \mathbf{X} \in \{0,1\}^{N\times N} \\ & \mathbf{X} \succeq \mathbf{0} \\ & \operatorname{rank}(\mathbf{X}) \le r \end{aligned} \tag{4}$$

The formulation in (3) involves maximizing a nonconvex quadratic objective function over a set of binary matrices $\mathbf{Z} \in \{0,1\}^{N\times r}$. One way to sidestep this difficulty would be by "lifting" the formulation from a low-dimensional space of $N \times r$ matrices to a high-dimensional space of $N \times N$ matrices by defining an auxiliary semidefinite matrix $\mathbf{X} = \mathbf{Z}\mathbf{Z}^\top$ that represents the clustering matrix and expressing the feasible space in terms of the valid inequalities for $\mathbf{X}$. The resulting formulation is given in (4). In the following proposition, we show that these two formulations are equivalent.

Proposition 1. *Formulations* (3) *and* (4) *are equivalent up to a rotation; that is, if* $\mathbf{X}^*$ *is an optimal solution to optimization problem* (4)*, then there exists a decomposition* $\mathbf{X}^* = \mathbf{G}^*\mathbf{G}^{*\top}$ *and an orthogonal matrix* $\mathbf{O} \in \mathbb{R}^{r\times r}$ *such that* $\mathbf{Z}^* = \mathbf{G}^*\mathbf{O}$ *is an optimal solution for* (3) *with the same objective function value.*

We defer the proof to the online appendix. Note that in the formulation presented in (4), the rows of $\mathbf{X}$ corresponding to outliers are essentially zero vectors. This provides us with a way to identify the outliers. However, even this formulation is a nonconvex optimization problem due to the rank and integrality constraints imposed on $\mathbf{X}$. Hence, we obtain tractable reformulations by considering two convex relaxations for the problem. In the first, we relax the binary constraint on $\mathbf{X}$ and also drop the rank constraint. This yields the following SDP formulation:

$$\begin{aligned} \underset{\mathbf{X}}{\text{maximize}} \quad & \langle \mathbf{K} - \gamma \mathbf{E}_N, \mathbf{X} \rangle \\ \text{subject to} \quad & 0 \le X_{ij} \le 1 \quad \forall i,j \\ & \mathbf{X} \succeq \mathbf{0}. \end{aligned} \tag{Robust-SDP}$$

We note here that similar SDP formulations have also been proposed in the community detection literature (Cai and Li 2015, Guédon and Vershynin 2016, Amini et al. 2018). Next, we consider a second relaxation in which we also allow the SDP constraint to be dropped from the formulation. The resulting formulation is a linear program that is specified below:

$$\begin{aligned} \underset{\mathbf{X}}{\text{maximize}} \quad & \langle \mathbf{K} - \gamma \mathbf{E}_N, \mathbf{X} \rangle \\ \text{subject to} \quad & 0 \le X_{ij} \le 1 \quad \forall i,j \end{aligned} \tag{Robust-LP}$$

Algorithm 1 (Robust Spectral Clustering/Robust-SDP)

Input: Observations $\mathbf{y}_1,\ldots,\mathbf{y}_N \in \mathbb{R}^d$, number of clusters r, scaling parameter $\theta \in \mathbb{R}_+$,and offset parameter $\gamma \in (0, 1)$.

1. Construct Gaussian kernel matrix $\mathbf{K}$, where $K_{ij} = \exp\left(\frac{-\|\mathbf{y}_i - \mathbf{y}_j\|^2}{\theta^2}\right)$.

2. Solve Robust-LP (Robust-SDP) to obtain the estimated clustering matrix $\hat{\mathbf{X}}$ ($\hat{\mathbf{X}}^{\text{SDP}}$).

3. Compute the top r eigenvectors of $\hat{\mathbf{X}}$ ($\hat{\mathbf{X}}^{\mathrm{SDP}}$) obtains $\hat{\mathbf{U}} \in \mathbb{R}^{N\times r}$.

4. Apply k-means clustering on rows of $\hat{\mathbf{U}}$ to estimate the cluster membership matrix $\hat{\mathbf{Z}}$.

5. Use $\hat{\mathbf{X}}$ ($\hat{\mathbf{X}}^{\mathrm{SDP}}$) to determine the degree threshold τ. Set $\hat{\mathcal{I}} = \{i \in [N] : \deg(i) \geq \tau\}$ and $\hat{\mathcal{O}} = [N]\backslash\hat{\mathcal{I}}$.

For convenience, we denote the feasible region of Robust-LP by set $\mathcal{X}$ and its optimal solution by $\hat{\mathbf{X}}$. It is straightforward to see that $\hat{\mathbf{X}}$ admits a simple analytical solution, which can be expressed below:

$$\hat{X}_{ij} = \begin{cases} 1 & \text{if } K_{ij} - \gamma > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Algorithm 1 summarizes the robust spectral clustering algorithm. To obtain the SDP variant of the algorithm, in step 2 of the algorithm, we solve the Robust-SDP formulation instead of the Robust-LP formulation. We also note here that steps 3 and 4 of the algorithm simply correspond to the application of vanilla spectral clustering to $\hat{\mathbf{X}}$. In general, solving the k-means clustering problem in step 4 is an NP-hard problem. Therefore, in our analysis, instead of solving the problem exactly, similarly to Lei and Rinaldo (2015), we consider the use of a $(1+\epsilon)$-approximate k-means clustering algorithm that runs in polynomial time in the number of datapoints n (Kumar et al. 2004). In the last step, we estimate the set of outliers $\hat{\mathcal{O}}$. Based on our derivations of the Robust-SDP and Robust-LP formulations, we note that the outlier points in $\mathcal{O}$ correspond to near-zero degree nodes in the true clustering matrix $\mathbf{X}^0$. We make use of this fact to determine a degree threshold τ from the degree distribution of the nodes in $\hat{\mathbf{X}}$ and assign the nodes that have degrees lesser than τ in $\hat{\mathbf{X}}$ to the set of outliers $\hat{\mathcal{O}}$. The main idea behind this procedure is that if $\hat{\mathbf{X}}$ closely approximates $\mathbf{X}^0$ and the threshold τ is appropriately chosen, then the low-degree nodes below the threshold in $\hat{\mathbf{X}}$ are good candidates for being outliers.

It is important to note that properly choosing the parameters θ and γ is central to the performance of the algorithm. For instance, if we choose the value of γ to be arbitrarily close to 0 or 1, then $\hat{\mathbf{X}}$ obtained after rounding is either an all ones matrix or an all zeros matrix, thereby rendering the denoising step useless. In Section 4, we derive theoretical values for θ and γ in terms of $\sigma_{\max}$ and $\Delta_{\min}$.

4. Main Results

In this section, we summarize our main results and provide an overview of the approach used to obtain these results. Our main theoretical result is a finite sample guarantee on the estimation error for $\hat{\mathbf{X}}$. Specifically, we show that the relative estimation error for $\hat{\mathbf{X}}$ decays exponentially in the square of the signal-to-noise ratio with probability tending to one as $N \to \infty$, provided there is sufficient separation between cluster centers and the number of outliers m are a small fraction of the number of inliers points n (Theorem 1). Using the result, we show that, provided the clusters are approximately balanced, the error rate for $\hat{\mathbf{X}}$ translates into an error rate for $\hat{\mathbf{Z}}$, and hence, the fraction of misclassified data points per cluster also decays exponentially in the square of the signal-to-noise ratio (Theorem 2).

For analyzing semidefinite relaxations of clustering problems, a rather useful direction is the approach described in Guédon and Vershynin (2016), which is in the context of stochastic block models. The main idea in the analysis of Guédon and Vershynin (2016) and Mixon et al. (2017) is to come up with a suitable reference matrix $\mathbf{R}$ and then use concentration of measure to control the deviation of the input matrix (adjacency matrix $\mathbf{A}$ for Guédon and Vershynin 2016, the matrix of pairwise squared Euclidean distances in Mixon et al. 2017, and the kernel matrix $\mathbf{K}$ for us) from the reference matrix. However, there are some important differences between our setting and theirs. SGMMs and SBMs are fundamentally different because the kernel matrix $\mathbf{K}$ constructed for a SGMM arises from n i.i.d. datapoints, leading to entries that are statistically dependent on each other. In contrast, the adjacency matrix of a random graph for a SBM has $\binom{n}{2}$ Bernoulli random variables, which are conditionally independent given the latent cluster memberships. Therefore, the analytical techniques required to analyze SGMMs are completely different compared with SBMs. Both Mixon et al. (2017) and Yan and Sarkar (2021) use suitable reference matrices for related but different SDP relaxations. The proof techniques that we develop in this section are new and involve coming up with a new reference matrix that allows us to carefully bound the tail probabilities. In addition, the resulting error bound that we get from our analysis is also tighter than that of the aforementioned papers.

We now provide an overview of our proof approach. Our constructed reference matrix $\mathbf{R} \in [0,1]^{N\times N}$ satisfies two properties:

(i) $\mathbf{R}$ is close to $\mathbf{K}$ with high probability in the ℓ_1-norm sense.

(ii) The solution to the reference optimization problem (5) defined below corresponds to the true clustering matrix $\mathbf{X}^0$ (Lemma 1).

$$\begin{aligned} \underset{\mathbf{X}}{\text{maximize}} \quad & \langle \mathbf{R} - \gamma \mathbf{E}_N, \mathbf{X} \rangle \\ \text{subject to} \quad & 0 \leq X_{ij} \leq 1 \quad \forall i,j \end{aligned} \tag{5}$$

In other words, the reference matrix $\mathbf{R}$ is chosen in a way such that the true clustering matrix $\mathbf{X}^0$ solves the reference optimization problem, which is obtained by replacing kernel matrix $\mathbf{K}$ in Robust-LP with $\mathbf{R}$.

We show that if (i) holds, then with high probability $\hat{\mathbf{X}} \in \mathcal{X}$ approximately solves the reference optimization problem in (5), that is, $\langle \mathbf{R} - \gamma \mathbf{E}_N, \hat{\mathbf{X}} \rangle \approx \langle \mathbf{R} - \gamma \mathbf{E}_N, \mathbf{X}^0 \rangle$ (see Lemma 3). Using this result, we then prove that if (ii) holds and the number of outliers is a small fraction of the number of inliers in the data set, then the estimated clustering matrix $\hat{\mathbf{X}}$ is close to the true clustering matrix $\mathbf{X}^0$. In other words, the relative estimation error is $\frac{\|\hat{\mathbf{X}} - \mathbf{X}^0\|_1}{\|\mathbf{X}^0\|_1} \leq \epsilon$ (small), with probability tending to one as $N \to \infty$ (see Theorem 1). Next, using the Davis-Kahan theorem (Yu et al. 2014), we show that, provided the clusters are relatively balanced in sizes, the error rates obtained for $\hat{\mathbf{X}}$ also hold for the clustering membership matrix $\hat{\mathbf{Z}}$ obtained by applying spectral clustering on $\hat{\mathbf{X}}$ (see Theorem 2).

For our analysis, we assume the reference matrix $\mathbf{R}$ to be a random matrix whose (i, j)-th entry is defined as below:

$$R_{ij} = \begin{cases} \max\{K_{ij}, \tau_{\text{in}}\} & \text{if both } i \text{ and } j \in \mathcal{C}_k \\ \min\{K_{ij}, \tau_{\text{out}}^{(k,l)}\} & \text{if } i \in \mathcal{C}_k, j \in \mathcal{C}_l \ (l \neq k) \\ \gamma & \text{if either } i \in \mathcal{O} \text{ or } j \in \mathcal{O} \end{cases} \tag{6}$$

Here, $\tau_{\text{in}} := \exp\left(-\frac{r_{\text{in}}^2}{\theta^2}\right)$ and $\tau_{\text{out}}^{(k,l)} := \exp\left(-\frac{r_{\text{out}}^{(k,l)2}}{\theta^2}\right)$ are threshold quantities defined respectively for the diagonal and off-diagonal blocks of reference matrix over the set of inlier points. For $i, j \in \mathcal{C}_k$, we obtain R_{ij} by thresholding K_{ij} to τ_{in} if $K_{ij} < \tau_{\text{in}}$. Similarly, for any $i \in \mathcal{C}_k$ and $j \in \mathcal{C}_l$, R_{ij} thresholds the value to $\tau_{\text{out}}^{(k,l)}$ if $K_{ij} > \tau_{\text{out}}^{(k,l)}$. The values of parameters r_{in} and $r_{\text{out}}^{(k,l)}$, which we specify later in the section, are determined such that with high probability only a few kernel entries violate the thresholds defined for their respective blocks, and thus property (i) is satisfied.

To ensure that our constructed reference matrix $\mathbf{R}$ satisfies property (ii), we impose a *strong assortativity* condition (similar to the analysis used for SBMs) that assumes that for the set of inlier points the smallest entry $R_{\text{min}}^{\text{in}}$ on the diagonal blocks of $\mathbf{R}$ is strictly greater than the largest entry $R_{\text{max}}^{\text{out}}$ on any of its off-diagonal blocks, that is,

$$R_{\text{min}}^{\text{in}} = \min_{i, j \in \mathcal{C}_k : k \in [r]} R_{ij} > \max_{i \in \mathcal{C}_k, j \in \mathcal{C}_l : k, l \in [r]} R_{ij} = R_{\text{max}}^{\text{out}}. \tag{7}$$

Based on the definition of the reference matrix, it is clear that $R_{\text{min}}^{\text{in}} \geq \tau_{\text{in}}$ and $R_{\text{max}}^{\text{out}} \leq \tau_{\text{out}} := \max_{k \neq l} \tau_{\text{out}}^{(k,l)}$. Thus, the strong assortativity condition in (7) is immediately implied if we require that $\tau_{\text{in}} > \tau_{\text{out}}$. We now use the strong assortativity condition in (7) to show that the true clustering matrix $\mathbf{X}^0$ is the solution to the reference optimization problem in (5) as required by property (ii).

Lemma 1. *Suppose that the strong assortativity condition in (7) holds and $R_{\text{max}}^{\text{out}} < \gamma < R_{\text{min}}^{\text{in}}$; then, the true clustering matrix $\mathbf{X}^0$ maximizes the reference optimization problem in (5).*

Proof. Set $R_{\text{max}}^{\text{out}} < \gamma < R_{\text{min}}^{\text{in}}$. Then, for the set of inlier points, all entries on the diagonal blocks of $\mathbf{R} - \gamma \mathbf{E}_N$ are strictly positive, whereas those on the off-diagonal blocks are strictly negative. Thus, $\mathbf{X}^0 = \arg\max_{\mathbf{X} \in [0,1]^{N \times N}} \langle \mathbf{R} - \gamma \mathbf{E}_N, \mathbf{X} \rangle$; that is, $\mathbf{X}^0$ maximizes the reference objective function over the feasible region comprising of all $[0,1]^{N \times N}$ matrices. □

Remark 1. Note that although we do not have SDP constraints, $\mathbf{X}^0 = \mathbf{Z}^0 \mathbf{Z}^{0\top} \in \mathcal{S}_N^+$, which implies that $\mathbf{X}^0 \in \mathcal{X}$ and $\mathbf{X}^0 \in \arg\max_{\mathbf{X} \in \mathcal{X}} \langle \mathbf{R} - \gamma \mathbf{E}_N, \mathbf{X} \rangle$. And thus, Lemma 1 also applies to Robust-SDP.

Next, we present Lemma 2, which provides a bound on the estimation error for the inlier parts of $\mathbf{X}^0$ and $\hat{\mathbf{X}}$ in terms of the difference in their corresponding objective function values for the reference optimization problem.

Lemma 2. *Suppose that the strong assortativity condition in (7) holds and $R_{\text{max}}^{\text{out}} < \gamma < R_{\text{min}}^{\text{in}}$; then, the estimation error for $\mathbf{X}^0$ over the set of inlier data points is*

$$\|\hat{\mathbf{X}}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}^0\|_1 \leq \frac{\langle \mathbf{R} - \gamma \mathbf{E}_N, \mathbf{X}^0 - \hat{\mathbf{X}} \rangle}{\min(R_{\text{min}}^{\text{in}} - \gamma, \gamma - R_{\text{max}}^{\text{out}})}.$$

Additionally, if the penalty parameter $\gamma \in (R_{\text{max}}^{\text{out}}, R_{\text{min}}^{\text{in}})$ is expressed as $\gamma = \upsilon \tau_{\text{in}} + (1 - \upsilon)\tau_{\text{out}}$ for some constant $\upsilon \in (0, 1)$, then the above bound simplifies to

$$\|\hat{\mathbf{X}}_{\mathcal{I}} - \mathbf{X}_{\mathcal{I}}^0\|_1 \leq \frac{\langle \mathbf{R} - \gamma \mathbf{E}_N, \mathbf{X}^0 - \hat{\mathbf{X}} \rangle}{\min\{\upsilon, 1 - \upsilon\}(\tau_{\text{in}} - \tau_{\text{out}})}.$$

In the next lemma, we show that if the kernel matrix is close to the reference matrix in a ℓ_1-norm sense, then the difference in the objective values of the reference optimization problem is also small.

Lemma 3. *Let $\mathbf{K}_{\mathcal{I}}, \mathbf{R}_{\mathcal{I}} \in [0,1]^{n \times n}$ denote respectively the parts of the kernel and reference matrices with each (i, j)-th entry restricted to the set of inlier points, and then*

$$\langle \mathbf{R} - \gamma \mathbf{E}_N, \mathbf{X}^0 - \hat{\mathbf{X}} \rangle \leq 2\|\mathbf{K}_{\mathcal{I}} - \mathbf{R}_{\mathcal{I}}\|_1.$$

Based on the definition of the reference matrix in (6), we note that for the (i, j)-th entry on the diagonal block of reference matrix where both $i, j \in \mathcal{C}_k$, R_{ij} deviates from its corresponding kernel value K_{ij} only if K_{ij} is below the threshold value τ_{in}. Similarly, for the (i, j)-th entry on the off-diagonal block where $i \in \mathcal{C}_k$ and $j \in \mathcal{C}_l$,

R_{ij} differs from K_{ij} only if K_{ij} is above the threshold value $\tau_{\text{out}}^{(k,l)}$ for that block. Therefore, we obtain a bound on $\|\mathbf{K}_{\mathcal{I}} - \mathbf{R}_{\mathcal{I}}\|_1$ by bounding the number of kernel entries that deviate from their respective threshold values on the diagonal and off-diagonal blocks. In particular, we can bound the ℓ_1-loss in Lemma 3 by the following:

$$2 \cdot \underbrace{\sum_{k\in[r]} \sum_{i,j\in\mathcal{C}_k : i<j} \mathbb{1}_{\{K_{ij}<\tau_{\text{in}}\}}}_{A} + \underbrace{\sum_{k\neq l} \sum_{i\in\mathcal{C}_k, j\in\mathcal{C}_l} \mathbb{1}_{\{K_{ij}>\tau_{\text{out}}^{(k,l)}\}}}_{B} \tag{8}$$

If the entries of the kernel matrix were independent, a straightforward application of standard concentration inequalities would have provided us a bound. However, because of the dependence between them, we use properties of the concept of U-statistics (Hoeffding 1963). In particular, we write the first part (A) of the above decomposition in terms of the following sum of one-sample U-statistics:

$$A = \sum_k \binom{n_k}{2} U_{kk}, \qquad U_{kk} = \frac{\sum_{\{(i,j):i,j\in\mathcal{C}_k, i<j\}} \mathbb{1}_{\{K_{ij}<\tau_{\text{in}}\}}}{n_k(n_k-1)/2}. \tag{9}$$

Similarly, we write the second part (B) of the decomposition in terms of the following sum of two-sample U-statistics:

$$B = \sum_{k\neq l} n_k n_l U_{kl}, \qquad U_{kl} = \frac{\sum_{i\in\mathcal{C}_k, j\in\mathcal{C}_l} \mathbb{1}_{\{K_{ij}>\tau_{\text{out}}^{(k,l)}\}}}{n_k n_l}. \tag{10}$$

A U-statistic of degree m is an unbiased estimator of some unknown quantity $\mathbb{E}[h(w_1,\ldots,w_m)]$ (where $w_1, \ldots, w_n$ are i.i.d. observations drawn from some underlying probability distribution). It can be written as an average of the h function (also known as the kernel function) applied on $\binom{n}{m}$ size m subsets of the data. It is not hard to see that U_{kk} defined in (9) is a U-statistic created from $\mathbf{y}_i, i\in\mathcal{C}_k$, where $\mathbf{y}_i$ are drawn i.i.d. from the k-th SGMM mixture component. On the other hand, U_{kl} defined in (10) is a two-sample U-statistic created from two i.i.d. data sets drawn from the k-th and l-th SGMM mixture component. Finally, using concentration results for U-statistics from Hoeffding (1963) and Arcones (1995), we obtain a probabilistic bound on the number of corrupt entries. This leads to the bound on the estimation error for $\hat{\mathbf{X}}$ in Theorem 1, which we present in the next subsection.

4.1. Estimation Error

We are now in a position to present our first main result, which states that if the number of outlier points is much smaller than the number of inlier points in the data set, then with probability tending to one, the error rate obtained is small, provided that there is enough separation between the cluster centers and the sample size is sufficiently large. We state this result formally in the theorem below.

Theorem 1 (Estimation Error for Robust-LP Solution $\hat{\mathbf{X}}$). *Let* $\tau_{\text{in}} = \exp\left(-\frac{5\Delta_{\min}^2}{32\theta^2}\right)$ *and* $\tau_{\text{out}}^{(k,l)} = \exp\left(-\frac{\Delta_{kl}^2}{2\theta^2}\right)$. *Choose* $\gamma \in (\tau_{\text{out}}, \tau_{\text{in}})$, *where* $\tau_{\text{out}} := \max_{k\neq l} \tau_{\text{out}}^{(k,l)} = \exp\left(-\frac{\Delta_{\min}^2}{2\theta^2}\right)$. *Suppose* $\theta = \Theta(\Delta_{\min})$ *and the minimum separation between cluster centers* $\Delta_{\min} \geq 8\sigma_{\max}\sqrt{d}$, and *then with probability at least* $1 - 2r/n_{\min}$, *we have that the estimation error for the inlier part of* $\hat{\mathbf{X}}$ *is*

$$\|\hat{\mathbf{X}}_{\mathcal{I}} - \mathbf{X}^0_{\mathcal{I}}\|_1 \leq Cn^2 \cdot \max\left\{\exp\left(-\frac{\Delta_{\min}^2}{64\sigma_{\max}^2}\right), \frac{\log n_{\min}}{n_{\min}}\right\}. \tag{11}$$

In addition, the relative estimation error for $\hat{\mathbf{X}}$ *is*

$$\frac{\|\hat{\mathbf{X}} - \mathbf{X}^0\|_1}{\|\mathbf{X}^0\|_1} \leq C'r\exp\left(-\frac{\Delta_{\min}^2}{64\sigma_{\max}^2}\right) + C''r\max\left\{\frac{\log n_{\min}}{n_{\min}}, \frac{m}{n}\right\}. \tag{12}$$

Here, $C, C', C'' > 0$ *are universal constants, and* $n_{\min} := \min_{k\in[r]} n_k > r$ *denotes the cardinality of the smallest cluster.*

Remark 2. In Section 4.3, we prove that if one does a suitable dimensionality reduction to first project the data on the top $r-1$ principal components, then with probability tending to one, the projected data becomes a SGMM in a $r-1$ dimensional space with minimum cluster separation $\Delta_{\min}/2$ as N goes to ∞. As a result, the new separation condition for applying Algorithm 1 to this projected data set becomes

$$\Delta_{\min} \geq 16\sigma_{\max}\sqrt{\min\{d, r\}}. \tag{13}$$

Remark 3. In the supplementary material (Theorem E.1), we show that for a mixture of Gaussians with identical covariance matrices, the separation condition can be further reduced to $d^{1/4}$ up to a logarithmic factor in SNR (which, in conjunction with the same argument as in Remark 2 gives a separation of $\min\{d,r\}^{1/4}$) to get an error rate polynomially decaying in the SNR.

From Theorem 1, we have that if there are no outliers in the data set, that is, $m = 0$, or if the number of outliers grow at a considerably slower rate compared with the number of inlier points, that is, $m = o_P(n)$, then asymptotically the error rate for $\hat{\mathbf{X}}$ decays exponentially with the square of the signal-to-noise ratio. To analyze this result in terms of prior theoretical work that has been done in the context of sub-Gaussian mixture models without any outliers, we note that Mixon et al. (2017) showed that for the

k-means clustering SDP proposed by Peng and Wei (2007), which assumes that the number of clusters r is known, the estimation error (obtained after rescaling) in a Frobenius norm sense $\|\hat{\mathbf{X}} - \mathbf{X}^0\|_F^2$ decays at a rate of $\frac{r^2 n_{\max}^2}{\mathrm{SNR}^2}$ provided the minimum separation $\Delta_{\min} \gtrsim r\sigma_{\max}$. In more recent work, Fei and Chen (2018) showed that for their SDP formulation that minimizes the k-means objective assuming all clusters to be equal-sized, the relative estimation error decays exponentially in the square of the signal-to-noise ratio provided $\Delta_{\min} \gtrsim \sqrt{r}\sigma_{\max}$. Giraud and Verzelen (2018) obtained a similar error rate for the k-means clustering SDP proposed by Peng and Wei (2007) that did not assume clusters to be equal-sized. Similarly to Fei and Chen (2018) and Giraud and Verzelen (2018), our result in Theorem 1 also guarantees a theoretical error bound that decays as $\exp(-\Omega(\mathrm{SNR}^2))$. The obtained bound is strictly better compared with Mixon et al. (2017), as shown below:

$$\|\hat{\mathbf{X}} - \mathbf{X}^0\|_F^2 \le \|\hat{\mathbf{X}} - \mathbf{X}^0\|_1 \lesssim n^2 \exp(-\Omega(\mathrm{SNR}^2)).$$

A key point to note in our results is that, in contrast to Fei and Chen (2018) and Mixon et al. (2017), our proof does not assume any prior knowledge about the number and sizes of clusters. In addition, Theorem 1 generalizes the analysis to incorporate outliers in the mixture of sub-Gaussians setting. However, the separation condition $\Delta_{\min} \gtrsim \sqrt{d}\sigma_{\max}$ does not generalize well to high-dimensional settings where $d \gg r$. To overcome this, later in this section, we propose a simple dimensionality reduction procedure that allows us to obtain the error rate in (12) for a reduced separation of $\Delta_{\min} \gtrsim \sqrt{\min\{r, d\}}\sigma_{\max}$ when r is known.

Very recently, Löffler et al. (2021) obtained an exponentially decaying bound in the square of the signal-to-noise ratio for the spectral clustering algorithm proposed by Vempala and Wang (2004). However, for their analysis, they assumed the data to be generated from a mixture of spherical Gaussians with identity covariance matrices. Furthermore, for their result to hold with high probability, the minimum separation $\Delta_{\min}$ needs to go to infinity. Based on the simple example considered in Figure 1, we also note that this algorithm is not robust to outliers.

We conclude this subsection with a comment on outliers. In our analysis so far, we have not made any specific assumptions on the distribution of the outlier points. However, one may have stronger theoretical results if such assumptions can be made; in particular, the following discussion shows that our algorithm can in fact tolerate $O(n)$ outlier points under suitable assumptions.

Remark 4. Based on the distance of each outlier point to its closest cluster center, we divide the set of outlier points into two sets consisting of "good" and "bad" outlier points. Intuitively, the "good" outlier points are far away from all the clusters, whereas the "bad" outlier points may be arbitrarily close to one or more clusters. It can be easily shown that any outlier point that is "bad" and close to a cluster center can potentially have as many as $\Omega\left(\frac{n}{r}\right)$ neighbors with high probability. For this reason, the first assumption that we make about the outlier points requires that the cardinality of the set of bad outlier points is at most $o(n)$. On the other hand, if the outlier points are good, that is, if they are far away from the clusters, then the set of good outlier points is potentially allowed to have a cardinality of $O(n)$. However, these good outlier points must either be isolated points or occur in small "bunches" or clusters so that the cardinality of any one cluster, comprising entirely of outlier points, is not too large (of the order $\Omega\left(\frac{n}{r}\right)$). One can ensure this by restricting the number of outlier points within a small neighborhood of each good outlier $i \in \mathcal{O}_g$ to $o(n)$.

We now mathematically formalize these notions.

Definition 1. We denote the set of good outlier points by $\mathcal{O}_g := \{i \in \mathcal{O} : \min_{k\in[r]} \|\mathbf{y}_i - \boldsymbol{\mu}_k\| \ge \sqrt{2}\Delta_{\min}\}$, which consists of outlier points whose distance from their closest cluster centers is at least above the threshold $\sqrt{2}\Delta_{\min}$. In addition, we also assume that for all $i \in \mathcal{O}_g$, the set of outlier neighboring points $\mathcal{N}_{\mathcal{O}}(i) := \left\{j \in \mathcal{O} : \|\mathbf{y}_i - \mathbf{y}_j\| \le \frac{\Delta_{\min}}{\sqrt{2}}\right\}$ has cardinality $o(n)$.

We summarize our main result in the proposition below.

Proposition 2. *Let $\mathcal{O}$ denote the set of outlier points. Let $\mathcal{O}_g \subset \mathcal{O}$ be good outliers satisfying Definition 1. Let $\mathcal{O}_b := \mathcal{O} \setminus \mathcal{O}_g$. Suppose the parameters γ and θ are chosen as described in Theorem 1 and the minimum separation between cluster centers $\Delta_{\min} \ge 8\sigma_{\max}\sqrt{d}$; then, provided that the size of $\mathcal{O}_g$ is $O(n)$, with probability at least $1 - 3r/n_{\min}$, we have that the relative estimation error for $\hat{\mathbf{X}}$ is*

$$\frac{\|\hat{\mathbf{X}} - \mathbf{X}^0\|_1}{\|\mathbf{X}^0\|_1} \le C' r \cdot \max\left\{\exp\left(-\frac{\Delta_{\min}^2}{64\sigma_{\max}^2}\right), \sqrt{\frac{\log n_{\min}}{n_{\min}}}\right\} + \frac{2r|\mathcal{O}_b|}{n} \tag{14}$$

Here, $C' > 0$ is a universal constant, and $n_{\min} := \min_{k\in[r]} n_k > r$ denotes the cardinality of the smallest cluster.

The proof of the theorem is deferred to the online appendix.

Remark 5. In Proposition 2, we have $|\mathcal{O}_g| = O(n)$, and as long as $|\mathcal{O}_b|/n$ is smaller than the first term, we have the same asymptotic rate as Theorem 1.

4.2. Rounding Error

As detailed in Algorithm 1, we recover cluster labels $\hat{\mathbf{Z}}$ from the estimated clustering matrix $\hat{\mathbf{X}}$ by applying spectral clustering on the columns of $\hat{\mathbf{X}}$. Our proof technique for analyzing the spectral clustering step is inspired by the approach discussed in Lei and Rinaldo (2015), where the authors relied on a $(1+\epsilon)$-approximate k-means clustering algorithm (Kumar et al. 2004) to cluster the rows of the matrix $\hat{\mathbf{U}} \in \mathbb{R}^{N\times r}$, whose columns consist of the r principal eigenvectors of $\hat{\mathbf{X}}$ that correspond to an embedding of each point in r-dimensional space. In the next theorem, we derive theoretical guarantees on the misclassification rate for the solution $\hat{\mathbf{Z}}$ obtained from this rounding procedure.

Theorem 2 (Clustering Error for Rounded Solution $\hat{\mathbf{Z}}$). *Let $\hat{\mathbf{Z}}$ be the estimated cluster membership matrix obtained by applying spectral clustering on $\hat{\mathbf{X}}$ using a $(1+\epsilon)$-approximate k-means clustering algorithm. Define $\bar{\epsilon}$ to denote the bound on the relative estimation error of $\hat{\mathbf{X}}$ in the right hand side of (12). Suppose $\frac{64(2+\epsilon)\bar{\epsilon}}{n_{\min}^2}\frac{n^2}{r} \le 1$ and the separation condition $\Delta_{\min} \ge 8\sigma_{\max}\sqrt{d}$ hold; then, with probability at least $1-2r/n_{\min}$, the cardinality of the set of misclassified data points $\mathcal{S}_k \subset \mathcal{C}_k$ for each $k \in [r]$ is upper bounded as*

$$\sum_{k\in[r]} \frac{|\mathcal{S}_k|}{n_k} \le 64(2+\epsilon)\frac{\|\mathbf{X}^0 - \hat{\mathbf{X}}\|_1}{n_{\min}^2}, \tag{15}$$

where $n_{\min} := \min_{k\in[r]} n_k > r$ denotes the cardinality of the smallest cluster.

Remark 6. Based on our discussion in Remark 2, if we adopt the dimensionality reduction procedure described in Section 4.3 to first project the data on the top $r-1$ principal components, and then the new separation condition for Theorem 2 to hold for the projected data set becomes Equation (13) as before.

We note that the added condition on $\bar{\epsilon}$ is required to translate the error of $\hat{\mathbf{X}}$ to misclassification error and is easily satisfied. If the clusters are balanced, that is, $n_{\min} = \Theta(n/r)$, then it will be satisfied as long as $\mathrm{SNR} = \Omega(\log r)$, n is large, and m/n is small. It can also be satisfied for an unbalanced setting at the expense of a larger SNR and large enough $n_{\min}$. Thus, from (15), we see that the average misclassification rate per cluster for inlier data points decays exponentially in the signal-to-noise ratio as well as N tends to infinity, provided that the clusters are balanced and m/n is sufficiently small. In our proof, we first analyze the approximate k-means clustering step and show that the average fraction of misclassified data points per cluster is upper bounded by $\|\hat{\mathbf{U}} - \mathbf{U}^0\mathbf{O}\|_{\mathrm{F}}$, where $\mathbf{U}^0 \in \mathbb{R}^{N\times r}$ represents the r principal eigenvectors of $\mathbf{X}^0$ and $\mathbf{O} \in \mathbb{R}^{r\times r}$ is the optimal rotation matrix. Next, using the Davis-Kahan theorem (Yu et al. 2014), we obtain a bound on the deviation $\|\hat{\mathbf{U}} - \mathbf{U}^0\mathbf{O}\|_{\mathrm{F}}$ in terms of $\|\mathbf{X}^0 - \hat{\mathbf{X}}\|_1$.

Remark 7. Based on the minimax results obtained in Lu and Zhou (2016), we note that for the SGMM setting in which there are no outliers, that is, $m = 0$, the error rate derived in (15) is optimal up to a constant factor in the exponent. Specifically, in Lu and Zhou (2016), the optimal rate has a factor of $1/8$ within the exponent as opposed to the $1/64$ factor that we obtain from (12) and (15). In Online Appendix H, we show that by narrowing down the range of values that γ can take, the $1/64$ factor in (12) can be reduced to $1/33$ to obtain a tighter bound.

Remark 8. It is easy to show that with minor modifications, the results in Theorems 1 and 2 also hold respectively for the solutions $\hat{\mathbf{X}}^{\mathrm{SDP}}$ and $\hat{\mathbf{Z}}^{\mathrm{SDP}}$ obtained from the Robust-SDP formulation.

4.3. Dimensionality Reduction for Large d

In this section, we extend our analysis to high-dimensional problems where $d \gg r$. Without loss of generality, we make the assumption that the inlier part of the data (data matrix excluding the outlier points) is centered at the origin, that is, mean $\boldsymbol{\mu} = \sum_{k\in[r]}\pi_k\boldsymbol{\mu}_k = 0$ for the sub-Gaussian mixture model. Under this assumption, because the r mean vectors can lie in at most $r-1$ dimensional space, we apply Algorithm 1 after dimensionality reduction. This is similar to previous works of Vempala and Wang (2004) on Gaussian mixture models. In order to maintain the independence of data points, similar to Chaudhuri et al. (2009) and Yan and Sarkar (2021), we split the data into two random parts. One part is used to compute the directions of maximum variance using principal component analysis (PCA) on its covariance matrix. The data points in the other part are projected along these principal directions to obtain their representations in a low-dimensional space.

In this procedure, we first randomly split the data matrix $\mathbf{Y}$ into two disjoint sets P_2 and P_1 with their respective cardinalities N_2 and $N_1 := N - N_2$. Using the points in P_2, we construct the sample covariance matrix $\hat{\boldsymbol{\Sigma}}_2 = \frac{\sum_{i\in P_2}(\mathbf{y}_i - \bar{\mathbf{y}}_2)(\mathbf{y}_i - \bar{\mathbf{y}}_2)^\top}{N_2}$, where $\bar{\mathbf{y}}_2 = \frac{\sum_{i\in P_2}\mathbf{y}_i}{N_2}$ and obtain the matrix $\mathbf{V}_{r-1}^{(2)} \in \mathbb{R}^{d\times(r-1)}$, whose columns consist of the top $r-1$ eigenvectors of $\hat{\boldsymbol{\Sigma}}_2$ that represent the $r-1$ principal components. We obtain the projection $\mathbf{y}_i'$ of each data point $i \in P_1$ by projecting $\mathbf{y}_i$ onto

the subspace spanned by the top $r-1$ eigenvectors of $\hat{\mathbf{\Sigma}}_2$, that is, $\mathbf{y}_i' = \mathbf{V}_{r-1}^{(2)\top}\mathbf{y}_i$. Sample splitting ensures that the projection matrix is independent of the data matrix that is being projected. Hence, the projected data points $\mathbf{y}_i'$ in the split P_1 of data set are independent of each other. This ensures that the key assumption of independence of data points that underlies Theorems 1 and 2 is satisfied.

Next, we show that, provided the number of outliers is small in comparison with the number of inlier data points, the original pairwise distances between cluster centers are largely preserved with high probability after projection. We state this result formally in the proposition below. In our result, we assume that the r cluster means span the $r-1$ dimensional space.

Proposition 3. *Assume that* $\sum_k \pi_k \boldsymbol{\mu}_k = 0$ *and* $N_2 = N^\alpha$ *for some* $0 < \alpha < 1$. *Let* $\mathbf{Y}^{\mathcal{O}} \in \mathbb{R}^{m\times d}$ *denote the outlier part of the data matrix and* $\mathbf{H} := \sum_k \pi_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^\top$ *such that its smallest positive eigenvalue* $\eta_{r-1}(\mathbf{H}) > 5\left(\sigma_{\max}^2 + C_1\sqrt{\frac{2\alpha d N^{1-\alpha}\log N}{n}} + C_2\left(\frac{m}{N} + \sqrt{\frac{\alpha \log N}{N^\alpha}}\right)\max\left\{\Delta_{\max}^2, \|\mathbf{Y}^{\mathcal{O}}\|_{2,\infty}^2\right\}\right)$ *for some universal constants* C_1 *and* C_2. *Then, the projections* $\mathbf{y}_i'$ *obtained for inlier data points in* P_1 *are independent sub-Gaussians in* $r-1$ *dimensional space. In addition, suppose* $\Delta_{\min}$ *denotes the minimum separation between any pair of cluster centers in the original d-dimensional space; then, the minimum separation after projection in the reduced space is* $\Delta_{\min}/2$, *with probability at least* $1 - \tilde{O}(r^2 N^{-\alpha})$.

The proof can be found in the online appendix. The condition on η_{r-1} essentially lower bounds the separation between the cluster means. For a simple symmetric equal-sized two-component mixture model, it is easy to see that η_{r-1} is proportional to the square of the distance between the cluster centers. It is important to note here that the sample splitting procedure discussed in this section is mainly for theoretical convenience to ensure that the projected data points are obtained independently of each other; in practice, as discussed in Chaudhuri et al. (2009), this step is usually not required. We note that the cardinality of set P_2 is a $N^{-(1-\alpha)}$ fraction of the total number of points in $\mathbf{Y}$, and hence, it vanishes for large N. On the other hand, the misclassification rate for our algorithm for the balanced clusters setting is upper bounded as $\sum_{k\in[r]} \frac{|\mathcal{S}_k|}{n_k} \lesssim Cr^2 \exp\left(-\frac{\Delta_{\min}^2}{64\sigma_{\max}^2}\right) + C'\frac{mr}{n}$, which is asymptotically nonvanishing. Therefore, the asymptotic error rate remains unaffected by sample splitting. If we make α very large, for example, using $N_2 = N/\log N$, then the condition on the smallest eigenvalue is less restrictive, but we only label $N(1-1/\log N)$ data points.

4.4. Extension to Weakly Separated Clusters

In this section, we consider the problem setup in which not all clusters have a minimum separation of $\Delta_{\min} = 8\sigma_{\max}\sqrt{d}$ between them, which is the condition required in Theorem 1 for the results to hold. Specifically, we extend the theoretical results obtained in Theorems 1 and 2 to show that if the separation between a pair of clusters is small, then with probability tending to one, it is possible to recover the "weakly separated" clusters as a single merged cluster with low error rate.

To achieve this, we define the threshold on the minimum separation to be $\Delta_0 := 8\sigma_{\max}\sqrt{d}$. We classify each cluster pair (k, l) as "weakly" or "well" separated based on whether $\Delta_{kl} < \Delta_0$ or $\Delta_{kl} \geq \Delta_0$, respectively. Let $\mathcal{S}_{\text{we}} := \{(k,l) : \Delta_{kl} < \Delta_0 \text{ for } k,l \in [r]\}$ denote the set of all weakly separated pair of cluster pairs, and then we redefine the reference matrix to incorporate for weakly separated clusters as below:

$$R_{ij} = \begin{cases} \max\left\{K_{ij}, \exp\left(-\frac{r_{\text{in}}^2}{\theta^2}\right)\right\} & \text{if } i,j \in \mathcal{C}_k \text{ or if } i \in \mathcal{C}_k, j \in \mathcal{C}_l \\ & \text{with } (k,l) \in S_{\text{we}} \\ \min\left\{K_{ij}, \exp\left(-\frac{r_{\text{out}}^{kl\,2}}{\theta^2}\right)\right\} & \text{if } i \in \mathcal{C}_k, j \in \mathcal{C}_l \text{ with } (k,l) \in S_{\text{we}}^c \\ \gamma & \text{if either } i \in \mathcal{O} \text{ or } j \in \mathcal{O} \end{cases} \tag{16}$$

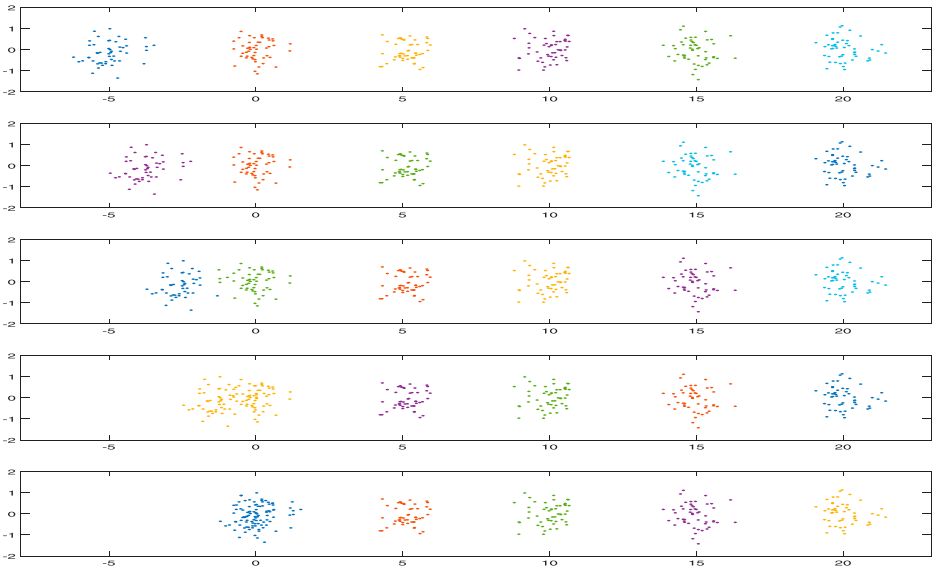
Clearly, if all clusters are well separated, the reference matrix defined above reduces to the reference matrix $\mathbf{R}$ in (6). However, under weak separation, we note that the solution $\tilde{\mathbf{X}}$ obtained from the reference optimization problem (5) corresponds to the solution where the weakly separated clusters form a single merged cluster and is of the form given below:

$$\tilde{X}_{ij} = \begin{cases} 1 & \text{if } i,j \in \mathcal{C}_k \text{ or if } i \in \mathcal{C}_k, j \in \mathcal{C}_l \text{ with } (k,l) \in S_{\text{we}} \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

Proposition 4. *Let* $\tilde{\mathbf{X}}$ *be the true solution defined in* (17) *and* $\hat{\mathbf{X}}$ *be the solution obtained from the Robust-LP formulation. Suppose* $\Delta' := \max_{k\neq l}\{\Delta_{kl} : \Delta_{kl} < \Delta_0\}$ *and* $\tilde{\Delta}_{\min} := \min_{k\neq l}\{\Delta_{kl} : \Delta_{kl} \geq \Delta_0\}$ *denote, respectively, the maximum cluster separation below threshold* Δ_0 *and the minimum cluster separation above* Δ_0. *Fix* $\gamma \in \left(\exp\left(\frac{-5\tilde{\Delta}_{\min}^2}{32\theta^2}\right), \exp\left(\frac{-\tilde{\Delta}_{\min}^2}{2\theta^2}\right)\right)$ *and set* $\theta = \Theta(\tilde{\Delta}_{\min})$. *Assume that* $\Delta' < \min\{\tilde{c}\tilde{\Delta}_{\min}, \Delta_0\}$, *and then with probability at least* $1 - 2r/n_{\min}$, *the estimation error for the inlier part of* $\hat{\mathbf{X}}$ *is upper bounded as*

$$\|\hat{\mathbf{X}}_{\mathcal{I}} - \tilde{\mathbf{X}}_{\mathcal{I}}\|_1 \leq Cn^2 \cdot \max\left\{\exp\left(-\frac{(\tilde{\Delta}_{\min} - \Delta'/\tilde{c})^2}{64\sigma_{\max}^2}\right), \frac{\log n_{\min}}{n_{\min}}\right\}. \tag{18}$$

Figure 2. (Color online) Example Shows the Effect of Reducing the Mean Cluster Separation Below the Threshold Δ_0



Notes. The original data set is obtained from a mixture of six spherical Gaussians with unit variances and a mean separation of 5 units. The separation between the first two clusters Δ_{12} is then incrementally reduced while keeping the separation between other clusters as fixed. Figure shows the final clustering obtained by applying the Robust-SC algorithm. As the overlap increases, the algorithm merges the first two clusters together.

In addition, the relative estimation error for $\hat{\mathbf{X}}$ *is*

$$\frac{\|\hat{\mathbf{X}}-\tilde{\mathbf{X}}\|_1}{\|\tilde{\mathbf{X}}\|_1} \le C' r \exp\left(-\frac{(\tilde{\Delta}_{\min}-\Delta'/\tilde{c})^2}{64\sigma^2_{\max}}\right) + C'' r \max\left\{\frac{\log n_{\min}}{n_{\min}}, \frac{m}{n}\right\}, \tag{19}$$

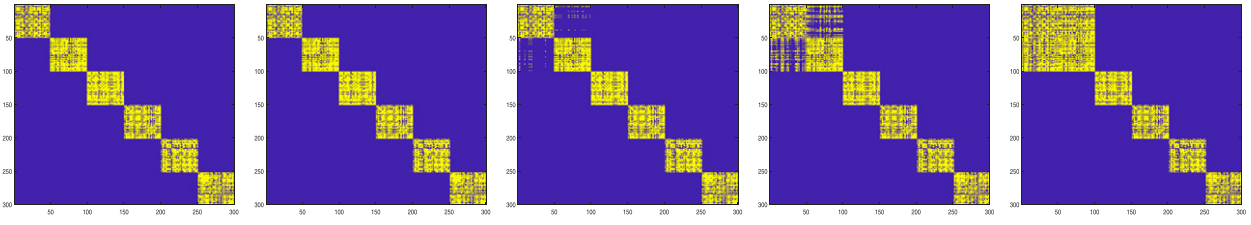
Here, C, C' *and* $\tilde{c}=\frac{\sqrt{10}}{8}$ *are positive constants.*

To understand the result, we consider a simple example (refer to Figure 2) where we have a mixture model consisting of six spherical Gaussians, with each having unit variance and a between-cluster separation of five units. We incrementally reduce the mean separation between the first two clusters Δ_{12} while keeping the separation between the remaining clusters as fixed. The clustering matrices $\hat{\mathbf{X}}$ obtained from the rounding step are shown in Figure 3. Because the mean separation between the first two clusters is decreased, we note that they get gradually merged in $\hat{\mathbf{X}}$, whereas the remaining part of $\hat{\mathbf{X}}$ corresponding to the "well" separated clusters remains unchanged. To obtain the final clustering of points from $\hat{\mathbf{X}}$, we first determine the number of clusters by adopting the procedure described in Section 5.6 based on the multiplicity of 0 eigenvalue(s) for the normalized graph Laplacian matrix. The corresponding clustering results obtained by applying the Robust-SC algorithm are shown in Figure 2.

5. Experiments

In this section, we study the performance of our Robust-LP-based spectral clustering algorithm (Robust-SC) on both simulated and real-world data sets. For our simulation studies, we conduct two different experiments. In the first experiment, we compare Robust-SC with three SDP-based clustering algorithms: (1) Robust-SDP, which is our proposed kernel clustering algorithm based on the Robust-SDP formulation; (2) Robust-Kmeans proposed by Kushagra et al. (2017), which is a regularized version of the *k*-means SDP formulation in Peng and Wei (2007); and (3) CC-Kmeans proposed by Rujeerapaiboon et al. (2019), which is another SDP-

Figure 3. (Color online) Clustering Matrices $\hat{\mathbf{X}}$ Obtained for Different Values of Δ_{12} Considered in the Example in Figure 2



Notes. As Δ_{12} is decreased, the overlap between the first two clusters in $\hat{\mathbf{X}}$ increases. However, the remaining part of $\hat{\mathbf{X}}$ remains unaffected.

based algorithm that recovers robust solutions by imposing explicit cardinality constraints for the clusters and the outlier points. Similar to Robust-SC and Robust-SDP algorithms, the formulations for both Robust-Kmeans and CC-Kmeans are capable of identifying outliers in datasets in addition to being robust to them. Therefore, we evaluate the performance of these algorithms in terms of both the inlier clustering accuracy and the outlier detection accuracy.

However, the SDP-based algorithms are computationally intensive to implement and, therefore, do not scale well to large-scale data sets. For this reason, in the second simulation experiment, we evaluate the performance of Robust-SC on larger data sets and compare it with three additional algorithms: (1) k-means++, (2) vanilla spectral clustering (SC), and (3) regularized spectral clustering (RegSC) (Joseph and Yu 2016, Zhang and Rohe 2018). Finally, for real-world data sets, we compare Robust-SC with all of the above-mentioned algorithms.

5.1. Implementation

We carried out all our experiments on a quadcore 1.9 GHz Intel Core i7-8650U CPU with 16GB RAM. For solving different SDP instances, we used the MATLAB package SDPNAL+ (Yang et al. 2015), which is based on an efficient implementation of a provably convergent ADMM-based algorithm.

5.2. Performance Metric

We measure the performance of algorithms in terms of clustering accuracy for the inliers and the percentage of outliers we can detect. We also report the overall accuracy, which is the total number of correctly clustered inliers and correctly detected outliers divided by N.

5.3. Parameter Selection

5.3.1. Choice of θ. It is well known that a proper choice of scaling parameter θ in the Gaussian kernel function plays a significant role in the performance of both spectral as well as SDP-based kernel clustering algorithms. We adopt the procedure prescribed by Shi et al. (2009) for choosing a good value of θ for low-dimensional problems. The main idea is to select θ in a way such that for $(1-\alpha)\times 100\%$ of the data points, at least a small fraction β (say around $5-10\%$) of the points in the neighborhood are within the "range" of the kernel function. In general, the value of selected β should be sufficiently high so that points that belong to the same cluster form a single component with relatively high similarity function values between them. Based on this idea, we choose θ as follows:

$$\theta = \frac{(1-\alpha)\text{ quantile of }\{q_1,\ldots,q_N\}}{\sqrt{(1-\alpha)\text{ quantile of }\chi^2_d}},$$

where for all points $1,\ldots,N$, each q_i equals the β quantile of the ℓ_2-distances $\{\|\mathbf{y}_i-\mathbf{y}_j\|, j=1,\ldots,N\}$ of point i from other points in the data set. Depending on the fraction of outlier points in the data set, we usually choose a small value of α so that for a majority of inlier points, the points in the neighborhood have a considerably higher similarity value. In all our experiments, we set $\beta=0.06$ and $\alpha=0.2$. For high-dimensional problems, we use the dimensionality reduction procedure described in Section 4 to first project the data points onto a low-dimensional space and then apply the above procedure to choose θ.

5.3.2. Choice of γ. Based on our discussion in Section 3, the parameter γ plays an equally important role in the performance of the Robust-LP formulation. For our experiments on simulated data sets, we choose the following value of γ:

$$\gamma = \exp\left(-\frac{t_\alpha}{2}\right),$$

where $t_\alpha = (1-\alpha)$ quantile of χ^2_d. This value is obtained by setting the distance in the Gaussian kernel function to equal the $(1-\alpha)$ quantile value of $\{q_1,\ldots,q_N\}$.

5.4. Simulation Studies

5.4.1. Comparison with SDP-based Algorithms. For the experiments in this section, we construct three synthetic data sets: (1) balanced spherical GMMs, (2) unbalanced spherical GMMs, and (3) balanced ellipsoidal GMMs. These data sets have been obtained from a mixture of linearly separable Gaussians and explore the effect of varying different model parameters like $\boldsymbol{\pi}$, $\{\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_r\}$ and $\{\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_r\}$ on the performance of the algorithms. In all of these data sets, we add outlier points in the form of uniformly distributed noise to the clusters. Table 2 lists out the model specifications for these synthetically generated data sets. Figure 4 depicts these data sets; in each part of this figure, the clusters formed by the inlier points are represented in different colors by solid circles, whereas the outlier points are marked with red x's.

As discussed earlier in this section, we compare the performance of our Robust-SC and Robust-SDP algorithms with two other SDP-based robust formulations, namely Robust-Kmeans and CC-Kmeans. In addition to explicitly requiring the number of outliers and cardinalities for all clusters as inputs, the CC-Kmeans algorithm suffers from several drawbacks. First, in contrast to both Robust-SDP and Robust-Kmeans, the algorithm requires solving the SDP formulation twice: once, to identify the outliers; and second, to recover the clusters after the outliers have been removed. Secondly, and more importantly, the CC-Kmeans formulation for r clusters, in

Table 2. Model Specifications for Synthetic Data Sets

Data set	Model specifications
1. Balanced Spherical GMMs	$\boldsymbol{\mu}_1 = [0,0]^\top, \boldsymbol{\mu}_2 = [6,3]^\top, \boldsymbol{\mu}_3 = [6,-3]^\top$ $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \text{Diag}([1,1])$ $n_1 = n_2 = n_3 = 150, m = 50$
2. Unbalanced Spherical GMMs	$\boldsymbol{\mu}_1 = [0,0]^\top, \boldsymbol{\mu}_2 = [20,3]^\top, \boldsymbol{\mu}_3 = [20,-3]^\top$ $\boldsymbol{\Sigma}_1 = \text{Diag}([5,5]), \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_3 = \text{Diag}([0.5,0.5])$ $n_1 = 500, n_2 = n_3 = 150, m = 50$
3. Balanced Ellipsoidal GMMs	$\boldsymbol{\mu}_1 = [0,5]^\top, \boldsymbol{\mu}_2 = [0,-5]^\top, \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \text{Diag}([20,1])$ $n_1 = n_2 = 200,\ m = 25$

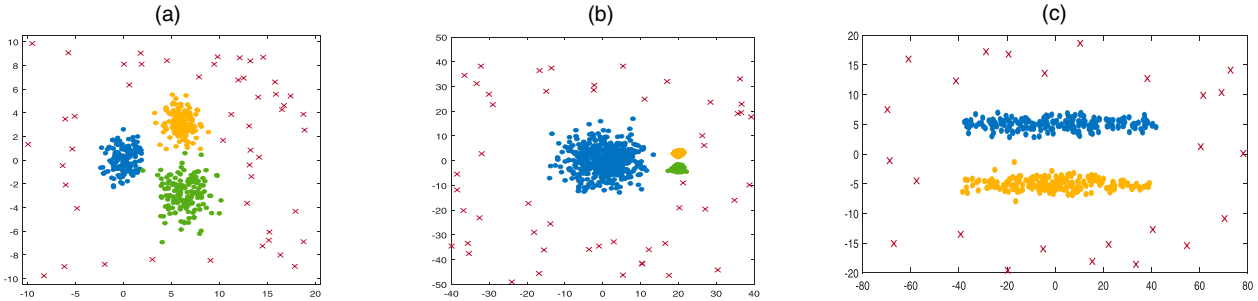
general, requires defining r separate matrix decision variables of dimensions $(N+1)\times(N+1)$, each with a positive semidefinite constraint. Because of extensive memory and computational requirements, the CC-Kmeans SDP could not be implemented on the synthetic data sets for the listed model specifications in Table 2. However, despite its several shortcomings, CC-Kmeans does provide us with a benchmark on the solution quality, provided the clustering problem has been entirely specified. Therefore, we try to evaluate the performance of CC-Kmeans algorithm by considering a smaller data set with a total of around $150-200$ data points in each data set, obtained by sampling an equal number of points from each cluster. We deliberately choose the clusters to be equal-sized for CC-Kmeans because when the clusters are equal-sized, the number of SDP variables per problem instance can be reduced (although each instance does need to be solved r times), thereby making the problem computationally tractable.

For each data set in Table 2, we generate 10 samples for the stated model specification and obtain clustering results for each algorithm except CC-Kmeans, for which we perform a single simulation run. Based on the implementation times in Table 3, it is quite evident that the CC-Kmeans algorithm is considerably slower (at least $10-20$ times) compared with the other SDP algorithms even for a down-sampled data set, and therefore, we do not show further experiments on CC-Kmeans in our simulation study.

We summarize the results obtained in Table 4. For each data set, we report the performance of the algorithms with respect to three metrics: (i) inlier clustering accuracy, (ii) outlier detection accuracy, and (iii) overall accuracy. On the balanced spherical GMMs data set, all of the algorithms perform equally well, with more than 95%($\pm$2%) overall accuracy. For the unbalanced spherical GMMs data set, Robust-SC and Robust-SDP are comparable, with about 98% ($\pm$0.6%) overall accuracy, whereas Robust-Kmeans performs poorly, with about 56%($\pm$2%) overall accuracy. Similarly, for the balanced ellipsoidal GMM data set, Robust-SC and Robust-SDP have similar accuracy values of 97.31%($\pm$0.6%) and 93.86%($\pm$5%), whereas Robust-Kmeans has a poor accuracy of 50.52%($\pm$1%).

Based on the high-accuracy values for inlier and outlier data points, Robust-SC and Robust-SDP consistently provide high-quality solutions in terms of recovering the true clusters for inlier data points as well as identifying outliers in the data set. On the other hand, whereas Robust-Kmeans and CC-Kmeans perform well for the balanced spherical GMMs data set, they fail either on the unbalanced spherical GMMs data set, where the clusters are unbalanced in terms of their cluster cardinalities (refer to Figure 5(a)), or the balanced ellipsoidal GMMs data set, where the clusters have significantly different variances along different directions (refer to Figure 5(b)).

Figure 4. (Color online) Synthetic Data Sets Generated for Evaluating the Performance of Clustering Algorithms



Note. (a) Balanced spherical GMMs; (b) unbalanced spherical GMMs; (c) balanced ellipsoidal GMMs.

Table 3. Solution Times (in Seconds) for Different Clustering Algorithms on Synthetic Data Sets

Data set	Robust-SC	Robust-SDP	Robust-Kmeans	CC-Kmeans
Balanced Spherical GMMs	3.24	265.62	355.65	3718
Unbalanced Spherical GMMs	3.18	828.56	1064.11	5726
Balanced Ellipsoidal GMMs	2.71	273.52	123.74	1944

Notes. For Robust-SC, Robust-SDP, and Robust-Kmeans, the solution times are specified for the entire data set, averaged over 10 simulation runs. For CC-Kmeans, the algorithm could not be implemented for the entire data set because of memory and computational limitations. Therefore, for comparison, we specify the run time for a single simulation on a down-sampled data set with an equal number of points from each cluster.

In addition, we note that although there is very little difference between Robust-SC and Robust-SDP in terms of solution quality, Robust-SC is orders of magnitude faster than Robust-SDP and other SDP-based algorithms in terms of solution times (refer to Table 3).

5.4.2. Comparison with *k*-Means++ and Spectral Clustering Algorithms. From the solution times reported in Table 3, it is quite evident that the SDP-based algorithms are intractable for large scale experiments. Therefore, in this section, we consider a much larger experiment setting and compare Robust-SC with more scalable *k*-means++ and spectral clustering algorithms.

In the experimental setup for this section, we assume that the n inlier points are generated in r-dimensional space from r equal-sized spherical Gaussians, which are centered at the vertices of a suitably scaled standard $(r-1)$-dimensional simplex and have identity covariance matrices. Thus, for all clusters, $k \in [r]$, $\boldsymbol{\mu}_k = s \cdot \mathbf{e}_k$ for some scale parameter s and $\boldsymbol{\Sigma}_k = \mathbf{I}_r$. The m outlier points are generated from another spherical Gaussian centered at the origin, that is, $\boldsymbol{\mu}_{\mathcal{O}} = \mathbf{0}$, and having a much larger variance ($\boldsymbol{\Sigma}_{\mathcal{O}} = 100 \cdot \mathbf{I}_r$).

We analyze the robustness of the Robust-SC algorithm under different model settings by varying the number of clusters (r), the number of outliers points (m), and the separation between cluster centers ($\Delta := \sqrt{2}s$). We compare Robust-SC with *k*-means++ and popular variants of the spectral clustering using clustering accuracy for inlier points as the evaluation metric. Figure 6 shows the results obtained. For this set of experiments, we assume that the default parameter values are set to $r = 15$, $s = 5$, $m = 400$, and $n/r = 400$. In each experiment, we assume that, except for the parameter that is varied, the other parameters are set to their default values. From the plots, we note that Robust-SC clearly outperforms the other clustering algorithms in terms of performance. We further demonstrate the scalability of the Robust-SC algorithm by repeating the experiment for $r = 50$ equal-sized clusters with $n = 50,000$ inlier points and $m = 1000$ outlier points. For 10 simulation runs of this experiment, we achieve an average inlier clustering accuracy of 0.9926 and an average solution time of 525.34 s with standard deviation values of 5.44×10^{-4} and 17.8 s respectively.

5.5. Real-World Data Sets

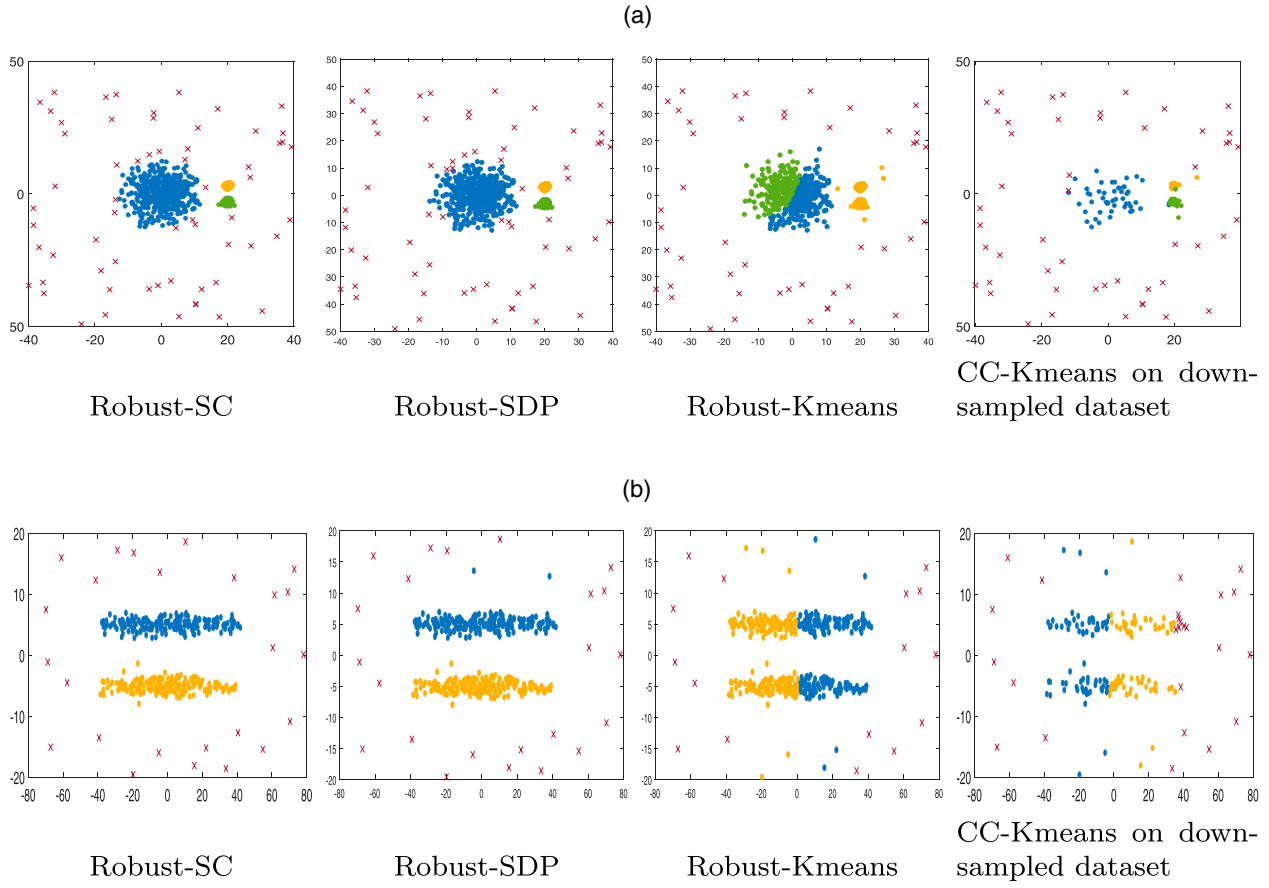
For evaluating the performance of different algorithms on real-world data sets, we standardize the data set by applying a z-score transformation to each attribute of the data set. For high-dimensional data sets, we adopt the dimensionality reduction procedure described in Section 4, which involves first computing the covariance matrix $\boldsymbol{\Sigma}$, projecting the data points onto the subspace spanned by the $r-1$ principal eigenvectors

Table 4. Performance of Clustering Algorithms on Synthetic Data Sets

Data set	Robust-SC		Robust-SDP		Robust-Kmeans		CC-Kmeans	
Balanced spherical	Inlier	0.9902	Inlier	0.9836	Inlier	0.9660	Inlier	1.0000
	Outlier	0.9840	Outlier	0.9080	Outlier	0.7540	Outlier	1.0000
	Overall	0.9896	Overall	0.9760	Overall	0.9448	Overall	1.0000
Unbalanced spherical	Inlier	0.9914	Inlier	0.9908	Inlier	0.5360	Inlier	0.9667
	Outlier	0.9680	Outlier	0.8840	Outlier	0.9240	Outlier	0.9600
	Overall	0.9900	Overall	0.9845	Overall	0.5588	Overall	0.9650
Balanced ellipsoidal	Inlier	0.9468	Inlier	0.9840	Inlier	0.5038	Inlier	0.4933
	Outlier	0.8080	Outlier	0.8000	Outlier	0.5280	Outlier	0.6800
	Overall	0.9386	Overall	0.9731	Overall	0.5052	Overall	0.5200

Notes. Performance of Robust-SC, Robust-SDP, and Robust-Kmeans algorithms in terms of their inlier clustering accuracy, outlier detection accuracy, and overall accuracy for synthetic data sets, averaged over 10 simulation runs. For CC-Kmeans, the algorithm could not be implemented for the entire data set because of memory and computational limitations. Therefore, for comparison, we specify the results for a single simulation on a down-sampled data set with an equal number of points from each cluster.

Figure 5. (Color online) Clustering Results for Different Algorithms on Synthetic Data Sets

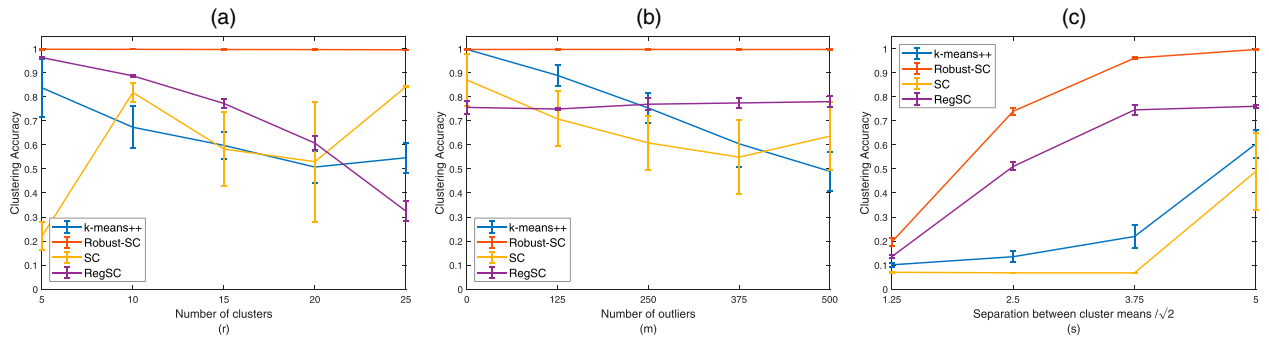


Notes. CC-Kmeans could not be implemented on the entire data set because of memory and computational limitations. Therefore, for comparison, we show the clustering results for a down-sampled data set with equal number of points from each cluster. (a) Unbalanced Spherical GMMs; (b) balanced ellipsoidal GMMs.

of $\mathbf{\Sigma}$ and then applying the z-score transformation to each attribute in the reduced space. All of these data sets were obtained from the UCI Machine Learning repository (Dua and Graff 2017). We provide below a brief description of these data sets and summarize their main characteristics in Table 5.

- **MNIST data set:** Handwritten digits data set comprised of 1,000 samples of 8×8 grayscale images (represented as a 64-dimensional vector) of digits from 0 to 9.
- **Iris data set:** Data set consists of a total of 150 samples from 3 clusters, each representing a particular type

Figure 6. (Color online) Effects of Varying the Model Parameters on the Inlier Clustering Accuracy for Different Algorithms



Notes. The default parameter values are set to $r = 15, s = 5, m = 400$ and $n/r = 400$. In each plot, apart from the parameter that is being varied, the other parameters are set to their default values.

Table 5. Real-World Data Sets with Their Main Characteristics

Data set	N - No. of data points	d - No. of dimensions	r - No. of clusters
MNIST	1000	64	10
Iris	150	4	3
USPS	500	256	4
Breast Cancer	683	9	2

of Iris plant. The four attributes associated with each data instance represent the sepal and petal lengths and widths of each flower in centimeters.

- **USPS data set:** A subset of the original USPS data set consisting of 500 random samples, each representing a 16×16 grayscale image of one of the following four digits: 0, 1, 3, and 7.
- **Breast cancer data set:** Data set consists of 683 samples of benign and malignant cancer cases. Every data instance is described by nine attributes, each having 10 integer-valued discrete levels.

For these real-world data sets, in addition to Robust-Kmeans and CC-Kmeans, we also compare the performances of Robust-SC and Robust-SDP with three other algorithms, namely k-means++, vanilla spectral clustering (SC), and regularized spectral clustering (RegSC).

As we discussed previously, for high-dimensional data sets, some form of a dimensionality reduction procedure is usually needed as an important preprocessing step. In the real-world data sets that we consider in our study, two data sets, namely MNIST and USPS, have high-dimensional features. Although none of the other methods that we compare our algorithm against explicitly recommends or analyzes the dimensionality reduction step for high-dimensional setting, for fairness, we apply our proposed dimensionality reduction procedure in Section 4 to all the algorithms. For reference, however, we consider a variant of the Robust-Kmeans algorithm, Robust-Kmeans-NoDR, that does not use our proposed dimensionality reduction procedure but is applied to the actual data in the original high-dimensional space.

Table 6 summarizes the clustering performance of different algorithms on the real-world datasets in terms of their overall accuracy for each data set. Based on the values in the table, we infer that both Robust-SC and Robust-SDP consistently perform well across all data sets and considerably better compared with the other algorithms considered in the study. Additionally, as we previously observed from our simulation studies, the Robust-SC algorithm recovers solutions that are almost as good as the Robust-SDP solutions and, for some data sets (MNIST and Breast Cancer), marginally better in terms of the clustering accuracy, even though Robust-SC is based on a simple rounding scheme, whereas the Robust-SDP algorithm requires solving the Robust-SDP formulation. For this reason, there is a significant disparity in the solution times noted for the two algorithms (refer to Figure 7), with the Robust-SC algorithm being approximately 100 times faster even for moderately sized problem instances. Additionally, comparing the performance of Robust-Kmeans and Robust-Kmeans-NoDR on the high-dimensional data sets, MNIST and USPS, we can easily see that the dimensionality reduction step significantly improves the performance of the algorithm on high-dimensional, real-world data sets.

5.6. Estimating Unknown Number of Clusters from Robust-SDP Formulation

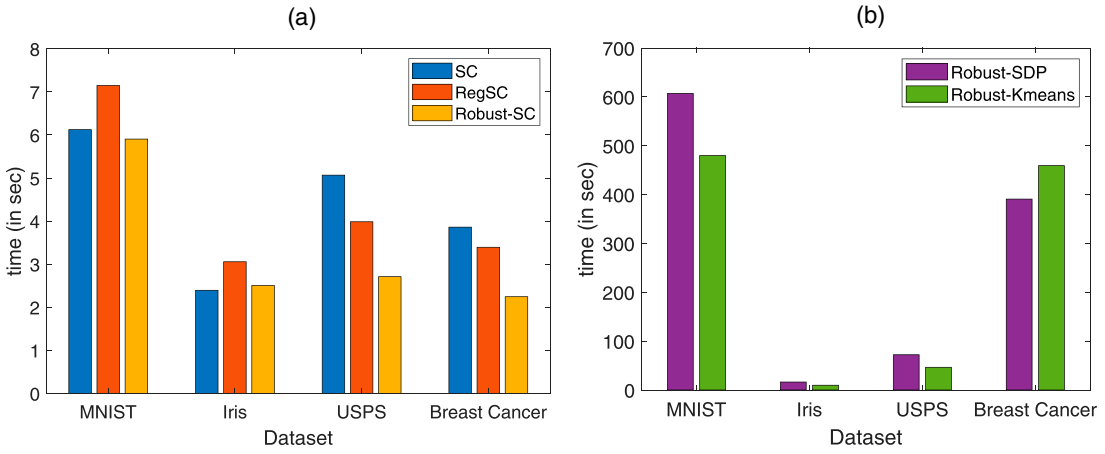
In several real-world problems, the number of clusters r is unknown. In this section, we discuss how we can obtain an estimate $\hat{r}$ for the number of clusters from the Robust-SDP solution $\hat{\mathbf{X}}^{\text{SDP}}$. In general, the SDP solution provides a more denoised representation of the kernel matrix as compared with the simple rounding scheme based on the Robust-LP solution. We propose a procedure based on the eigengap heuristic (Von Luxburg 2007) of the normalized graph Laplacian matrix $\mathbf{L}_{\tilde{\mathcal{I}}} := \mathbf{I} - \mathbf{D}_{\tilde{\mathcal{I}}}^{-1/2}\hat{\mathbf{X}}_{\tilde{\mathcal{I}}}^{\text{SDP}}\mathbf{D}_{\tilde{\mathcal{I}}}^{-1/2}$, where $\mathbf{D}_{\tilde{\mathcal{I}}} = \text{Diag}(\hat{\mathbf{X}}_{\tilde{\mathcal{I}}}^{\text{SDP}}\mathbf{1}_{|\tilde{\mathcal{I}}|})$ and $\tilde{\mathcal{I}} = \{i : \deg(i) \geq \tilde{\tau}\}$. Here, the threshold $\tilde{\tau}$ corresponds to some quantile $\tilde{\beta}$ of $\{\deg(i), i = 1, \ldots, N\}$. The key idea behind this heuristic is to select a value of $\hat{r}$ such that the $\hat{r}$ smallest eigenvalues $\lambda_1 \leq \ldots \leq \lambda_{\hat{r}}$ of $\mathbf{L}_{\tilde{\mathcal{I}}}$ are extremely small (close to 0), whereas $\lambda_{\hat{r}+1}$ is relatively large. The main argument for using the eigengap heuristic comes from matrix perturbation theory, which leverages the

Table 6. Performance of Different Clustering Algorithms on Real-World Data Sets

Algorithm	MNIST	Iris	USPS	Breast Cancer
Robust-SDP	0.8450	**0.8933**	**0.9720**	0.9649
Robust-SC	**0.8630**	0.8800	0.9620	**0.9722**
Robust-Kmeans	0.8040	0.8267	0.8320	0.9575
Robust-Kmeans-NoDR	0.6680	0.8267	0.6420	0.9575
CC-Kmeans	—	0.8400	—	—
SC	0.8580	0.6600	0.3280	0.6471
RegSC	0.7320	0.5200	0.6000	0.8873
k-means++	0.7850	0.8133	0.6080	0.9575

Notes. Performance of different clustering algorithms on real-world data sets in terms of their overall clustering accuracy. Entry with '-' indicates that the algorithm failed to terminate within the specified time limit of 2 hours.

Figure 7. (Color online) Solution Times (in Seconds) for Different Algorithms on Real-World Data Sets



Note. (a) Spectral methods; (b) SDP-based methods.

fact that if a graph consists of r disjoint clusters, then its graph Laplacian matrix has an eigenvalue of 0 with multiplicity r and its $(r+1)$-st smallest eigenvalue λ_{r+1} is comparatively larger.

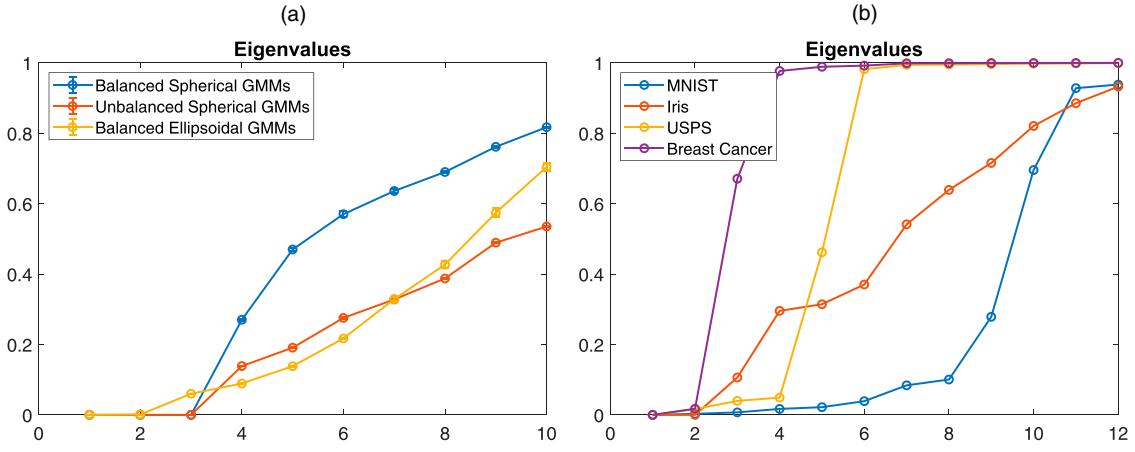
Figure 8 denotes the eigenvalues of the normalized graph Laplacian matrix for both synthetic and real-world data sets. From the plot, it is easy to see that the eigengap heuristic correctly predicts the number of clusters for each of the three synthetic data sets. It is important to note that the eigengap heuristic for finding the number of clusters usually works better when the signal-to-noise ratio is large, that is, either when the clusters are well separated or when the noise around the clusters is small. However, for many real-world data sets, a high signal-to-noise ratio is not always observed. For example, in the MNIST handwritten digits data set, there are considerable overlaps between clusters that represent digits 1 and 7 as well as digits 4 and 9. Thus, when the eigengap heuristic is applied on the MNIST data set, it returns $\hat{r} = 8$ as an estimate for the number of clusters. Similarly, for the iris data set, two of the clusters (Verginica and Versicolor) are known to intersect each other (Ana and Jain 2003). Thus, when the number of clusters is not specified, we get $\hat{r} = 2$ instead of the actual three clusters in the data set.

Although it is possible to obtain an estimate of r by applying the above procedure on the rounded matrix $\hat{\mathbf{X}}$ obtained from the Robust-LP formulation, we see that $\hat{r}$ obtained from $\hat{\mathbf{X}}^{\text{SDP}}$ is more accurate.

Acknowledgments

The authors are grateful to Rachel Ward for valuable comments on the paper and to the editors and reviewers for their constructive feedback.

Figure 8. (Color online) Eigenvalues of the Normalized Graph Laplacian Matrix $\mathbf{L}_{\tilde{\mathcal{I}}} := \mathbf{I} - \mathbf{D}_{\tilde{\mathcal{I}}}^{-1/2} \hat{\mathbf{X}}_{\tilde{\mathcal{I}}}^{\text{SDP}} \mathbf{D}_{\tilde{\mathcal{I}}}^{-1/2}$ for Synthetic and Real-World Data Sets with $\tilde{\beta} = 0.8$



Note. (a) Synthetic data sets; (b) real-world data sets.

References

Altman NS (1992) An introduction to kernel and nearest-neighbor nonparametric regression. *Amer. Statist.* 46(3):175–185.

Amini AA, Levina E (2018) On semidefinite relaxations for the block model. *Ann. Statist.* 46(1):149–179.

Amini AA, Razaee ZS (2021) Concentration of kernel matrices with application to kernel spectral clustering. *Ann. Statist.* 49(1):531–556.

Amini AA, Chen A, Bickel PJ, Levina E (2013) Pseudo-likelihood methods for community detection in large sparse networks. *Ann. Statist.* 41(4):2097–2122.

Ana L, Jain AK (2003) Robust data clustering. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Madison WI*, volume 2.

Arcones MA (1995) A Bernstein-type inequality for U-statistics and U-processes. *Statist. Probab. Lett.* 22(3):239–247.

Arora S, Kannan R (2001) Learning mixtures of arbitrary Gaussians. *Proc. 33rd Annual ACM Symposium on Theory of Computing*, 247–257 (ACM, New York).

Awasthi P, Sheffet O (2012) Improved spectral-norm bounds for clustering: Approximation, randomization, and combinatorial optimization. *Lecture Notes in Comput. Sci.*, 7408 (Springer, Heidelberg), 37–49.

Awasthi P, Bandeira AS, Charikar M, Krishnaswamy R, Villar S, Ward R (2015) Relax, no need to round: Integrality of clustering formulations. *Proc. 2015 Conf. Innovations in Theoretical Comput. Sci.*, 191–200 (ACM, New York).

Bickel PJ, Chen A, Levina E (2011) The method of moments and degree distributions for network models. *Ann. Statist.* 39(5): 2280–2301.

Bojchevski A, Matkovic Y, Günnemann S (2017) Robust spectral clustering for noisy data: Modeling sparse corruptions improves latent embeddings. *Proc. 23rd ACM SIGKDD Interat. Conf. Knowledge Discovery and Data Mining* (ACM, New York), 737–746.

Cai TT, Li X (2015) Robust and computationally feasible community detection in the presence of arbitrary outlier nodes. *Ann. Statist.* 43(3):1027–1059.

Chaudhuri K, Kakade SM, Livescu K, Sridharan K (2009) Multi-view clustering via canonical correlation analysis. *Proc. 26th Annual Internat. Conf. Machine Learn.*, 129–136.

Cover T, Hart P (1967) Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* 13(1):21–27.

Cuesta-Albertos JA, Gordaliza A, Matrán C (1997) Trimmed *k*-means: An attempt to robustify quantizers. *Ann. Statist.* 25(2):553–576.

Dasgupta S (1999) Learning mixtures of Gaussians. *40th Annual Sympos. on Foundations of Comput. Sci.* (IEEE), 634–644.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B.* 39(1):1–22.

Dhillon IS, Guan Y, Kulis B (2004) Kernel k-means: spectral clustering and normalized cuts. *Proc. Tenth ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (ACM, New York), 551–556.

Ding C, He X (2004) K-nearest-neighbor consistency in data clustering: incorporating local information into global optimization. *Proc. 2004 ACM Symposium on Applied Comput.* (ACM, New York), 584–589.

Du M, Ding S, Jia H (2016) Study on density peaks clustering based on k-nearest neighbors and principal component analysis. *Knowl. Base. Syst.* 99:135–145.

Dua D, Graff C (2017) UCI machine learning repository. http://archive.ics.uci.edu/ml.

El Karoui N (2010) On information plus noise kernel random matrices. *Ann. Statist.* 38(5):3191–3216.

Fei Y, Chen Y (2018) Hidden integrality of SDP relaxations for sub-Gaussian mixture models. *Proc. 31st Conf. on Learn. Theory*, vol. 75, 1931–1965.

Fishkind DE, Sussman DL, Tang M, Vogelstein JT, Priebe CE (2013) Consistent adjacency-spectral partitioning for the stochastic block model when the model parameters are unknown. *SIAM J. Matrix Anal. Appl.* 34(1):23–39.

Forero PA, Kekatos V, Giannakis GB (2012) Robust clustering using outlier-sparsity regularization. *IEEE Trans. Signal Process.* 60(8): 4163–4177.

Franti P, Virmajoki O, Hautamaki V (2006) Fast merative clustering using a k-nearest neighbor graph. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(11):1875–1881.

Giraud C, Verzelen N (2018) Partial recovery bounds for clustering with the relaxed *K*-means. *Math. Stat. Learning* 1(3):317–374.

Goemans MX, Williamson DP (1995) Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM.* 42(6):1115–1145.

Guédon O, Vershynin R (2016) Community detection in sparse networks via Grothendieck's inequality. *Probab. Theory Related Fields* 165(3-4):1025–1049.

Hastie T, Tibshirani R (1996) Discriminant adaptive nearest neighbor classification and regression. *Adv. Neural Inf. Process. Syst.* 8:409–415.

Heckel R, Bölcskei H (2015) Robust subspace clustering via thresholding. *IEEE Trans. Inform. Theory* 61(11):6320–6342.

Heckel R, Tschannen M, Bölcskei H (2017) Dimensionality-reduced subspace clustering. *Info. Inference J IMA* 6(3):246–283.

Hoeffding W (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58(301):13–30.

Holland PW, Laskey KB, Leinhardt S (1983) Stochastic blockmodels: First steps. *Soc. Networks.* 5(2):109–137.

Iguchi T, Mixon DG, Peterson J, Villar S (2015) On the tightness of an SDP relaxation of k-means. Preprint, submitted May 18, https://arxiv.org/abs/1505.04778.

Joseph A, Yu B (2016) Impact of regularization on spectral clustering. *Ann. Statist.* 44(4):1765–1791.

Kumar A, Kannan R (2010) Clustering with spectral norm and the k-means algorithm. *51st Annual IEEE Symposium on Foundations of Computer Science* (IEEE, New York), 299–308.

Kumar A, Sabharwal Y, Sen S (2004) A simple linear time (1 + ε)-approximation algorithm for k-means clustering in any dimensions. *45th Annual IEEE Symposium on Foundations of Computer Science* (IEEE, New York), 454–462.

Kushagra S, McNabb N, Yu Y, Ben-David S (2017) Provably noise-robust, regularised *k*-means clustering. Preprint, submitted November 30, https://arxiv.org/abs/1711.11247.

Le CM, Levina E, Vershynin R (2015) Sparse random graphs: regularization and concentration of the laplacian. Preprint submitted April 23, https://arxiv.org/abs/1502.03049.

Lei J, Rinaldo A (2015) Consistency of spectral clustering in stochastic block models. *Ann. Statist.* 43(1):215–237.

Löffler M, Zhang AY, Zhou HH (2021) Optimality of spectral clustering in the Gaussian mixture model. *Ann. Statist.* 49(5):2506–2530.

Li Z, Liu J, Chen S, Tang X (2007) Noise robust spectral clustering. *IEEE 11th International Conference on Computer Vision*, (IEEE, New York) 1–8.

Li X, Li Y, Ling S, Strohmer T, Wei K (2020) When do birds of a feather flock together? k-means, proximity, and conic programming. *Math. Program.* 179(1–2):295–341.

Lloyd S (1982) Least squares quantization in PCM. *IEEE Trans. Inform. Theory.* 28(2):129–137.

Lu Y, Zhou HH (2016) Statistical and computational guarantees of Lloyd's algorithm and its variants. Preprint, December 7, https://arxiv.org/abs/1612.02099.

McSherry F (2001) Spectral partitioning of random graphs. *Proc. 42nd IEEE Sympos. Foundations of Computer Science* (IEEE, New York), 529–537.

Mixon DG, Villar S, Ward R (2017) Clustering subgaussian mixtures by semidefinite programming. *Info. Inference J IMA* 6(4):389–415.

Montanari A, Sen S (2016) Semidefinite programs on sparse random graphs and their application to community detection. *Proc. 48th Annual Sympos. Theory Comput.* (ACM, New York), 814–827.

Newman ME (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, 103(23):8577–8582.

Ng AY, Jordan MI, Weiss Y (2002) On spectral clustering: Analysis and an algorithm. *Adv. Neural Inform. Processing Systems*, 849–856.

Pearson K (1936) Method of moments and method of maximum likelihood. *Biometrika* 28(1–2):34–59.

Peng J, Wei Y (2007) Approximating k-means-type clustering via semidefinite programming. *SIAM J. Optim.* 18(1):186–205.

Qin T, Rohe K (2013) Regularized spectral clustering under the degree-corrected stochastic blockmodel. *Adv. Neural Inform. Processing Systems*, 3120–3128.

Rohe K, Chatterjee S, Yu B (2011) Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* 39(4): 1878–1915.

Rujeerapaiboon N, Schindler K, Kuhn D, Wiesemann W (2019) Size matters: Cardinality-constrained clustering and outlier detection via conic optimization. *SIAM J. Optim.* 29(2):1211–1239.

Schiebinger G, Wainwright MJ, Yu B (2015) The geometry of kernelized spectral clustering. *Ann. Statist.* 43(2):819–846.

Shi J, Malik J (2000) Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22(8):888–905.

Shi T, Belkin M, Yu B (2009) Data spectroscopy: Eigenspaces of convolution operators and clustering. *Ann. Statist.* 37(6B):3960–3984.

Soltanolkotabi M, Candés EJ (2012) A geometric analysis of subspace clustering with outliers. *Ann. Statist.* 40(4):2195–2238.

Soltanolkotabi M, Elhamifar E, Candès EJ (2014) Robust subspace clustering. *Ann. Statist.* 42(2):669–699.

Sussman DL, Tang M, Fishkind DE, Priebe CE (2012) A consistent adjacency spectral embedding for stochastic blockmodel graphs. *J. Amer. Statist. Assoc.* 107(499):1119–1128.

Van der Vaart AW (2000) *Asymptotic Statistics*, vol. 3. (Cambridge University Press, Cambridge, UK).

Vempala S, Wang G (2004) A spectral algorithm for learning mixture models. *J. Comput. System Sci.* 68(4):841–860.

Verdinelli I, Wasserman L (2018) Analysis of a mode clustering diagram. *Electron. J. Stat.* 12(2):4288–4312.

Vershynin R (2012) Introduction to the non-asymptotic analysis of random matrices. Eldar Y, Kutyniok G, eds. *Compressed Sensing: Theory and Applications* (Cambridge University Press, Cambridge, UK), 210–268.

Vinayak RK, Hassibi B (2016) Similarity clustering in the presence of outliers: Exact recovery via convex program. *IEEE International Symposium on Information Theory (ISIT)*, 91–95.

Von Luxburg U (2007) A tutorial on spectral clustering. *Statistics and Computing*. 17(4):395–416.

Von Luxburg U, Belkin M, Bousquet O (2008) Consistency of spectral clustering. *Ann. Statist.* 36(2):555–586.

Wang YX, Xu H (2013) Noisy sparse subspace clustering. *Proc. 30th Internat. Conf. on Machine Learning*, 28:89–97.

Wang Y, Wang YX, Singh A (2018) A theoretical analysis of noisy sparse subspace clustering on dimensionality-reduced data. *IEEE Trans. Inform. Theory*, 65(2):685–706.

Yan B, Sarkar P (2016) On robustness of kernel clustering. *Adv. Neural Inform. Processing Systems*, 3098–3106.

Yan B, Sarkar P (2021) Covariate regularized community detection in sparse graphs. *J. Amer. Statist. Assoc.* 116(534):734–745.

Yan B, Sarkar P, Cheng X (2018) Provable estimation of the number of blocks in block models. Storkey AJ, Pérez-Cruz F, eds. *Internat. Conf. on Artificial Intelligence and Statistics, AISTATS 2018, Spain, Proc. Machine Learning Research*, 84:1185–1194 (PMLR).

Yang L, Sun D, Toh KC (2015) SDPNAL+: A majorized semismooth newton-CG augmented lagrangian method for semidefinite programming with nonnegative constraints. *Math. Program. Comput.* 7(3):331–366.

Yu SX, Shi J (2003) Multiclass spectral clustering. *Proc. 9th IEEE International Conference on Computer Vision, vol. 2*, (IEEE, New York), 313.

Yu Y, Wang T, Samworth RJ (2014) A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika* 102(2):315–323.

Zhang Y, Rohe K (2018) Understanding regularized spectral clustering via graph conductance. *Adv. Neural Inf. Process. Syst.* 31: 10631–10640.

Prateek R. Srivastava is an applied scientist on the Middle Mile Marketplace Science team at Amazon. Prior to joining Amazon, he completed his MS and PhD degrees in Operations Research and Industrial Engineering from the University of Texas at Austin.

Purnamrita Sarkar is an assistant professor of Statistics at the University of Texas at Austin. She holds a PhD degree from the Machine Learning department at Carnegie Mellon University. Her research interests are at the intersection of asymptotic statistics, scalable algorithms, networks, and resampling methods for networks.

Grani A. Hanasusanto is an assistant professor of Operations Research and Industrial Engineering at The University of Texas at Austin. He holds a PhD degree in Operations Research from Imperial College London and an MSc degree in Financial Engineering from the National University of Singapore. His research focuses on the design and analysis of tractable solution schemes for decision-making problems under uncertainty, with applications in operations management, energy systems, machine learning, and data analytics.