Eigen selection in spectral clustering: a theory guided practice

Xiao Han*, Xin Tong†, and Yingying Fan‡

Abstract

Based on a Gaussian mixture type model of K components, we derive eigen selection procedures that improve the usual spectral clustering algorithms in high-dimensional settings, which typically act on the top few eigenvectors of an affinity matrix (e.g., $\mathbf{X}^\top\mathbf{X}$) derived from the data matrix $\mathbf{X}$. Our selection principle formalizes two intuitions: (1) eigenvectors should be dropped when they have no clustering power; (2) some eigenvectors corresponding to smaller spiked eigenvalues should be dropped due to estimation inaccuracy. Our selection procedures lead to new spectral clustering algorithms: ESSC for $K = 2$ and GESSC for $K > 2$. The newly proposed algorithms enjoy better stability and compare favorably against canonical alternatives, as demonstrated in extensive simulation and multiple real data studies.

KEY WORDS: clustering, eigen selection, low-rank models, high dimensionality, asymptotic expansions, eigenvectors, eigenvalues.

1. INTRODUCTION

Clustering is a widely-used unsupervised learning approach to divide observations into subgroups without the guidance of labels. It is an obvious statistical and machine learning formulation when there are no meaningful labels in the datasets, such as in customer segmentation and criminal cyber-profiling applications. It is also a sensible approach when labels, in theory, do exist, but we have solid reasons to believe that the labels in the datasets are far from accurate. For instance, Medicare-Medicaid fraud detection cannot be formulated as a supervised learning problem, because

*Xiao Han, International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, Anhui, China, 230026(Email: xhan011@e.ntu.edu.sg).

†Xin Tong, Data Sciences and Operations Department, University of Southern California, Los Angeles, CA, 90089(Email: xint@marshall.usc.edu). Tong is the corresponding author.

‡Yingying Fan, Data Sciences and Operations Department, University of Southern California, Los Angeles, CA, 90089(Email: fanyingy@marshall.usc.edu).

This work is partially supported by NSF of China (No. 12001518).

although the labeled fraudulent transactions are real frauds, people believe that there are a large number of undiscovered frauds in the record.

Over the last sixty years, many clustering approaches have been proposed. The most dominant ones include k-means, hierarchical clustering, spectral clustering, and various variants [Hastie et al., 2009, James et al., 2014]. The k-means algorithms [Bradley et al., 1999, Witten and Tibshirani, 2010] adopt a centroid-based clustering approach. Hierarchical clustering algorithms [Ward Jr, 1963] first seek to build a hierarchy of clusters and then make a cut at a hierarchical level. Spectral clustering [Ng et al., 2002, Von Luxburg, 2007] clusters observations using the spectral information of some affinity matrix derived from the original data matrix $\mathbf{X}$ for measuring the similarity among observations.

Among the above mentioned main-stream clustering approaches, spectral clustering is particularly well suited for high-dimensional settings, which refers to the situations that the number of features is comparable to or larger than the sample size. High-dimensional settings mainly emerged with modern biotechnologies such as microarray and remain relevant due to the subsequent technological advances such as next-generation sequencing (NGS) technologies. Methodological and theoretical questions in high-dimensional supervised learning (i.e., regression and classification) have been attracting a great deal of attention in the statistics community over the last 20 years (see the review paper Zou [2019] and references within). In contrast, high-dimensional unsupervised problems have had far fewer works so far. It is a challenging problem mainly because effective dimension reduction is difficult without the assistance of a response variable. Spectral clustering alleviates the curse of dimensionality in high-dimensional clustering by consulting only a few less noisy eigenvectors of an affinity matrix. For example, suppose that we would like to cluster n observations into K groups, where K is the predetermined cluster number. Spectral clustering algorithms usually first compute the top few eigenvectors of an affinity matrix (e.g., $\mathbf{X}^\top\mathbf{X}$ which we adopt in this work) and then perform a k-means step using just these eigenvectors.

The rationale behind the above spectral clustering method is that under a broad data matrix generative model of low-rank mean matrix plus noise, the latent data label information is completely captured by the population eigenvectors corresponding to the top (i.e., spiked) eigenvalues of $\mathbb{E}\mathbf{X}^\top\mathbb{E}\mathbf{X}$, which can be estimated by the eigenvectors corresponding to the spiked eigenvalues

of the affinity matrix $\mathbf{X}^\top\mathbf{X}$. Thus, the eigenvectors corresponding to non-spiked eigenvalues can be safely dropped and the purpose of noise reduction is achieved. However, there are some potential issues in the common spectral clustering implementation. Taking $K = 2$ as an example. It could be that (1) one of the top two eigenvectors can be useless either because its coordinates are all equal or its corresponding population eigenvalue is 0; (2) even if both top eigenvectors are useful in clustering, the noise matrix may introduce too much noise to the second eigenvector so that it cannot be estimated accurately. Therefore, selecting the eigenvectors has a potential to improve spectral clustering methods. Simulation results summarized in Tables 1-2 support intuitions (1)-(2), respectively, under models A and B ($K = 2$ for both), whose exact settings can be found in Section S.1 in the Supplementary Material. Here the misclustering rate is defined as $\inf_{\pi\in\mathfrak{M}_K} \frac{1}{n}\sum_{i=1}^n \mathbf{1}(\pi(\widehat{Y}_i) = Y_i)$, where Y_i and $\widehat{Y}_i$ are the true label and estimated label respectively with $Y_i \in \mathfrak{N}_K$, $\mathfrak{N}_K$ is the label set containing K different entries and $\mathfrak{M}_K$ is the collection of all one-to-one and onto mappings from $\mathfrak{N}_K$ to itself. Loosely, model A has one (of top 2) eigenvector with no clustering power, and model B has both top eigenvectors with clustering power but the second one cannot be accurately estimated. These tables also suggest that the benefit of eigen selection is more profound with higher dimensionality.

Table 1: Misclustering rates with standard error in parentheses for Model A

methods \ p	100	200	400	600	800
drop the useless eigenvector before k-means	.01(.0006)	.014(.0006)	.02(.0008)	.029(.0008)	.04(.0015)
k-means on both eigenvectors	.01(.0005)	.014(.0006)	.028(.0004)	.064(.0108)	.097(.0099)

Table 2: Misclustering rates with standard error in parentheses for Model B

methods \ p	100	200	400	600	800
k-means on the first eigenvector	.01(.001)	.011(.0011)	.019(.0015)	.025(.0014)	.036(.0022)
k-means on both eigenvectors	.023(.0016)	.04(.003)	.118(.0092)	.182(.0104)	.251(.0124)

In this paper, we first formalize the above intuitions by considering the special case of $K = 2$ and Gaussian distributions. Concretely, the data matrix follows the aforementioned structure of low rank mean matrix plus noise defined as $\mathbf{X} = \mathbb{E}\mathbf{X} + (\mathbf{X} - \mathbb{E}\mathbf{X})$, where $\mathbf{X}$ is a $p \times n$ matrix and n is the sample size. A natural and popular way is to construct the affinity matrix as $\mathbf{X}^\top\mathbf{X}$ *. We show that the top two eigenvectors of $\mathbf{H} := (\mathbb{E}\mathbf{X})^\top\mathbb{E}\mathbf{X}$, which can be understood as the noiseless

*A comparison with one alternative affinity matrix construction is given in subsection 3.2

version of the affinity matrix, completely capture the label information. We also identify scenarios where exactly one of these two eigenvectors of $\mathbf{H}$ has clustering power. Note that the eigenvectors of $\mathbf{H}$ are unavailable to us and any operation has to be applied to their sample counterparts, that is, the eigenvectors of the affinity matrix $\mathbf{X}^\top\mathbf{X}$.

We propose an innovative **e**igen **s**election procedure in the usual **s**pectral **c**lustering algorithms and name the resulting algorithm ESSC for $K = 2$. Our eigenvector selection step is guided by the theoretical investigation of the top two eigenvectors of $\mathbf{H}$, and justified by analysis on sample-level eigen properties. Our theoretical development does not require a sparsity assumption on the data generative model, such as those in Cai et al. [2013] and Jin and Wang [2016]. This suggests that our procedure is potentially suitable for a wider range of applications. A by-product of our theoretical development is an asymptotic expansion of the eigenvalues when the population eigenvalues are close to each other (Proposition 1). This is a result of stand-alone interest.

The intuition of eigenvector selection in the case of multiple clusters ($K \geq 3$) is slightly different in the sense that a useful eigenvector may only have partial clustering power because some clusters can collapse along the direction of that eigenvector. Thus, the second intuition discussed above of dropping eigenvectors with less estimation accuracy is over-weighted by including all eigenvectors with clustering power. For this reason, we recommend to only practice the first intuition, i.e., screening out eigenvectors without any clustering power, which include those corresponding to the non-spiked eigenvalues and those whose coordinates are equal. Based on this, we propose a new algorithm GESSC, in which "G" stands for "generalized."

We provide extensive simulation studies, and observe that in a vast array of settings, the newly proposed clustering algorithms ESSC and GESSC compare favorably in terms of stability and mis-clustering rates against the spectral clustering algorithm without the eigen selection step. Although our theoretical analysis is conducted under a Gaussian distribution assumption, the general idea of eigenvector selection extends to other settings and other high-dimensional clustering problems such as community detection using network data.

Although the eigen selection idea for spectral clustering, minus the common practice of dropping eigenvectors corresponding to the non-spiked eigenvalues, is mostly absent in the statistics community, it was practiced in one previous work in the computer science literature. Indeed, Xiang and

Gong [2008] proposed an EM algorithm to select the eigenvectors of an affinity matrix. But their approach is a heuristic practice and lacks theoretical analysis for the eigenvalues and eigenvectors to support the method.

There is relatively recent literature on theoretical and methodological developments on high-dimensional clustering. For instance, Ng et al. [2002] proposed a symmetric-Laplacian-matrix-based spectral clustering approach and prove the corresponding consistency. Cai et al. [2019] proposed a clustering procedure based on the EM algorithm for a high-dimensional Gaussian mixture model and proved consistency and minimax optimality for the procedure. Jin and Wang [2016] proposed a Kolmogorov-Smirnov (KS) score based feature selection approach (IF-PCA) to first reduce the feature dimension before implementing spectral clustering on a centered version of the data. The feature selection idea for clustering was also considered in other works including Chan and Hall [2010] and Azizyan et al. [2013]. None of these aforementioned works select eigenvectors. In this sense, our method and theory complement the existing literature by providing a way to stabilize and improve the performance of existing spectral clustering methods.

The rest of the paper is organized as follows. We introduce the statistical model and key notations in Section 2. In Section 3, we present the main algorithm ESSC for $K = 2$ and detailed rationale that leads to it. Section 4 includes the theoretical results regarding ESSC. Section 5 introduces a multi-cluster (i.e., $K > 2$) extension GESSC together with its theoretical backing. Simulation study and real data analysis are conducted in Sections 6 and 7 respectively, followed by a short discussion. Technical lemmas, proofs and further discussion are relegated to the Supplementary Material.

2. MODEL SETTING AND NOTATIONS

In the methodological development and theoretical analysis, we consider the following sampling scheme. We assume that the data matrix $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is generated from

$$\mathbf{x}_i = Y_i\boldsymbol{\mu}_1 + (1 - Y_i)\boldsymbol{\mu}_2 + \mathbf{w}_i,\ i = 1, \ldots, n\,, \tag{1}$$

where $\{\mathbf{w}_i\}_{i=1}^n$ are i.i.d. from p-dimensional Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\mu}_2$ are two p-dimensional non-random vectors, and $Y_1, \ldots, Y_n \in \{0, 1\}$ are deterministic latent class labels. As

such, $Y_i = 1$ means that the ith observation $\mathbf{x}_i$ is from class 1, and $Y_i = 0$ means that $\mathbf{x}_i$ is from class 2. The parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}$ are assumed to be unknown. Without loss of generality, we assume that $\boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_2 \neq 0$. The main objective is to recover the latent labels Y_i's from the data matrix $\mathbf{X}$. If $\{Y_i\}_{i=1}^n$ were i.i.d Bernoulli random variables, model (1) would be a Gaussian mixture model. Our analysis can extend to that setting but we opt for considering fixed Y_i's to focus on our attention to the eigen selection principle.

We introduce some notations that will be used throughout the paper. For a matrix $\mathbf{B}$, denote by $\sigma_k(\mathbf{B})$ the k-th largest singular value of $\mathbf{B}$ and and by $\|\mathbf{B}\|$ its spectral norm. For any vector $\mathbf{x}$, $\mathbf{x}(i)$ represents the i-th coordinate of $\mathbf{x}$. For any random matrix (or vector) $\mathbf{A}$, we use $\mathbb{E}\mathbf{A}$ to denote its expectation. We define $c_{11} = \|\boldsymbol{\mu}_1\|_2^2$, $c_{22} = \|\boldsymbol{\mu}_2\|_2^2$ and $c_{12} = \boldsymbol{\mu}_1^\top \boldsymbol{\mu}_2$, where $\|\cdot\|_2$ is the L_2 norm of a vector. For any positive sequences u_n and v_n, if there exists some positive constant c such that $u_n \geq c v_n$ for all $n \in \mathbb{N}$, then we denote $u_n \gtrsim v_n$. We denote the i-th largest eigenvalue of a square matrix $\mathbf{A}$ by $\lambda_i(\mathbf{A})$. Finally, we denote $\sigma_n^2 = \|\boldsymbol{\Sigma}\|^2(n+p)$.

3. ESSC FOR $K = 2$

Based on Model 1, we develop a novel eigen selection procedure that improves the widely used spectral clustering algorithms. We start our reasoning from the noiseless case. The entire logic flow of the development process is presented before we introduce the final **e**igen-**s**elected **s**pectral **c**lustering algorithm (ESSC).

3.1 Motivation if the signal were known

When people believe there are two clusters, a common spectral clustering practice is to perform k-means on the top $K = 2$ eigenvectors of $\mathbf{X}^\top \mathbf{X}$. We offered in the introduction some intuitions and numerical evidence about why this might be improvable. We will formalize these intuitions in this section.

For notational convenience, denote $\mathbf{a}_1 = \mathbf{y} = (Y_1, \ldots, Y_n)^\top$ and $\mathbf{a}_2 = \mathbf{1} - \mathbf{y}$. Let $n_1 = \|\mathbf{a}_1\|_2^2$ and $n_2 = \|\mathbf{a}_2\|_2^2$, then n_1 and n_2 are the numbers of non-zero components of $\mathbf{a}_1$ and $\mathbf{a}_2$ respectively, and $n_1 + n_2 = n$. A noiseless counterpart of $\mathbf{X}^\top \mathbf{X}$ is $\mathbf{H} = (\mathbb{E}\mathbf{X})^\top \mathbb{E}\mathbf{X}$. By Model 1, $\mathbf{H}$ can be decomposed by

$$\mathbf{H} = \mathbf{a}_1 \mathbf{a}_1^\top c_{11} + \mathbf{a}_2 \mathbf{a}_2^\top c_{22} + \mathbf{a}_1 \mathbf{a}_2^\top c_{12} + \mathbf{a}_2 \mathbf{a}_1^\top c_{12} \geq 0\,. \tag{2}$$

Let d_1 and d_2 be the top two singular values of $\mathbb{E}\mathbf{X}$, and $\mathbf{u}_1$ and $\mathbf{u}_2$ be the corresponding right singular vectors. Then d_i^2 and $\mathbf{u}_i$ are the corresponding eigenvalue and eigenvector of $\mathbf{H}$, $i = 1, 2$. By the discussions in Section S.2 of the Supplementary Material, the eigenvector $\mathbf{u}$ corresponding to a nonzero eigenvalue $d^2 > 0$ takes *at most* two distinct values in its components. Moreover, if $d^2 > 0$ and $\mathbf{u}$ takes two distinct values in its components, then these values have a one-to-one correspondence with the cluster labels. We also notice that when $d^2 = 0$, $\mathbf{u}$ would not be informative for clustering. Given these observations, we introduce the following definition for ease of presentation.

Definition 1. *A population eigenvector* $\mathbf{u}$ *of* $\mathbf{H}$ *is said to have clustering power if its corresponding eigenvalue* d^2 *is positive and its coordinates take exactly two distinct values.*

Theorem 1. *The top two eigenvalues of* $\mathbf{H}$ *can be expressed as*

$$d_1^2 = \frac{1}{2}\left(n_1c_{11} + n_2c_{22} + (n_1^2c_{11}^2 + n_2^2c_{22}^2 + 4n_1n_2c_{12}^2 - 2n_1n_2c_{11}c_{22})^{\frac{1}{2}}\right), \tag{3}$$

and

$$d_2^2 = \frac{1}{2}\left(n_1c_{11} + n_2c_{22} - (n_1^2c_{11}^2 + n_2^2c_{22}^2 + 4n_1n_2c_{12}^2 - 2n_1n_2c_{11}c_{22})^{\frac{1}{2}}\right). \tag{4}$$

Moreover, we conclude the following regarding the clustering power of $\mathbf{u}_1$ *and* $\mathbf{u}_2$*.*

(a) When $c_{12}^2 = c_{11}c_{22}$*, the problem is degenerate with* $d_1^2 = n_1c_{11} + n_2c_{22}$ *and* $d_2^2 = 0$*, and only the eigenvector* $\mathbf{u}_1$ *has clustering power.*

(b) When $c_{12}^2 \neq c_{11}c_{22}$*,* $c_{12} = 0$ *and* $n_1c_{11} = n_2c_{22}$*, we face the problem of multiplicity (i.e.,* $d_1^2 = d_2^2 = n_1c_{11}$*) and at least one of* $\mathbf{u}_1$ *and* $\mathbf{u}_2$ *have clustering power.*

(c) When $c_{12}^2 \neq c_{11}c_{22}$*,* $c_{12} = 0$ *and* $n_1c_{11} \neq n_2c_{22}$*, we have* $d_1^2 = \max\{n_1c_{11}, n_2c_{22}\}$ *and* $d_2^2 = \min\{n_1c_{11}, n_2c_{22}\} > 0$*, and both* $\mathbf{u}_1$ *and* $\mathbf{u}_2$ *have clustering power.*

(d) When $c_{12}^2 \neq c_{11}c_{22}$ *and* $c_{12} \neq 0$*, if* $n_1c_{11} + n_2c_{12} = n_2c_{22} + n_1c_{12}$*, exactly one eigenvector has clustering power, and if* $n_1c_{11} + n_2c_{12} \neq n_2c_{22} + n_1c_{12}$*, both eigenvectors have clustering power.*

We note that similar results to (b)–(d) of Theorem 1 were proved for Degree Corrected Stochastic Block Model in Lemma 1.1 of Jin [2015], with the difference that Jin [2015] considered data matrix

with 0/1 values that are independent on and above the diagonals. Theorem 1 implies that under our model described in equation (1), at least one of $\mathbf{u}_1$ and $\mathbf{u}_2$ have clustering power. More importantly, this theorem indicates that even in the noiseless setting (i.e., when $\mathbf{H}$ is known), there are cases in which only one eigenvector has clustering power and that this eigenvector could be either $\mathbf{u}_1$ or $\mathbf{u}_2$. This suggests the potential importance of eigenvector selection in spectral clustering and we propose Oracle Procedure 1 below to select a set $\mathcal{U}$ of important eigenvectors under the noiseless setting.

Algorithm 1 [Oracle Procedure 1]

1: Set $\mathcal{U} = \emptyset$.
2: Check whether $\mathbf{u}_1$ has two distinct values in its components. If yes, add $\mathbf{u}_1$ to $\mathcal{U}$ and go to Step 3; If no, add $\mathbf{u}_2$ to $\mathcal{U}$ and go to Step 5.
3: Check whether $d_2^2 > 0$. If no, go to Step 5; If yes, go to Step 4.
4: Check whether $\mathbf{u}_2$ has two distinct values in its components. If yes, add $\mathbf{u}_2$ to $\mathcal{U}$ and go to Step 5; if no, go to Step 5.
5: Return $\mathcal{U}$.
6: Use the eigenvector(s) in $\mathcal{U}$ for clustering.

Despite its simple form, Oracle Procedure 1 is difficult to implement at the sample level. To elaborate, note that in practice we will have to estimate the eigenvalues and eigenvectors $(d_i^2, \mathbf{u}_i)$, $i = 1, 2$. Without loss of generality, assume that $d_1 \geq d_2 \geq 0$. Note that d_1 and d_2 are the top two singular values of $\mathbb{E}\mathbf{X}$, which can be naturally estimated by the top two singular values of $\mathbf{X}$. Further note that $\mathbf{u}_1$ and $\mathbf{u}_2$ are the top two right singular vectors of $\mathbb{E}\mathbf{X}$, which can be naturally estimated by $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$, the top two right singular vectors of $\mathbf{X}$, respectively. One useful technique in the literature for obtaining these sample estimates is to consider the linearization matrix

$$\mathcal{Z} = \begin{pmatrix} 0 & \mathbf{X}^\top \\ \mathbf{X} & 0 \end{pmatrix},$$

which is a symmetric random matrix with low-rank mean matrix. It can be shown that the top two singular values of $\mathbf{X}$ are the same as the top two eigenvalues of $\mathcal{Z}$, and the corresponding right singular vectors of $\mathbf{X}$, after rescaling, are subvectors of the top two eigenvectors of $\mathcal{Z}$.

It has been proved in the literature that for random matrices with expected low rank structure, such as $\mathcal{Z}$, the estimation accuracy of spiked eigenvectors largely depends on the magnitudes

of the corresponding eigenvalues. Specifically, as shown in Abbe et al. [2020+], the entrywise estimation error for each spiked eigenvector is of order inversely proportional to the magnitude of the corresponding eigenvalue. Thus, a dense eigenvector may be estimated very poorly unless the corresponding eigenvalue has a large magnitude (e.g., highly spiked). The results in Abbe et al. [2020+] apply to a large Gaussian ensemble matrix with independent entries on and above the diagonal. Similar conclusions can be found in Fan et al. [2020+] and Bao et al. [2020+] under Wigner or generalized Wigner matrix assumption.

Since spectral clustering is applied to estimated eigenvectors, the above-mentioned existing results suggest that in a high-dimensional two-class clustering, one should drop the second eigenvector in spectral clustering if the corresponding eigenvalue is not spiked enough, unless it is critical to include it, when, for example, the first spiked population eigenvector has no clustering power.

On the other extreme, if the two spiked eigenvalues are the same, that is, in the case of multiplicity, by part (b) of Theorem 1, at least one of $\mathbf{u}_1$ and $\mathbf{u}_2$ has clustering power. We argue that in this situation, at the sample level it is better to use both spiked eigenvectors in clustering for at least two reasons. *First*, by Proposition 1 to be presented in Section 4 and the remark after it, each d_i, $i = 1, 2$, can only be estimated with accuracy $O_p(1)$. Therefore, detecting the exact multiplicity can be challenging. *Second*, the two spiked population eigenvectors are not identifiable. The two spiked sample eigenvectors estimate some rotation of $(\mathbf{u}_1, \mathbf{u}_2)$, each with estimation accuracy of order inversely proportional to d_1 (or d_2) [Abbe et al., 2020+]. Thus, even in the worst case where exactly one eigenvector is useful, including both in clustering will not deteriorate the clustering result much because the additional estimation error caused by the useless eigenvector is in the same order as caused by the useful eigenvector. In view of the discussions above, we update the oracle procedure as follows. Our implementable algorithm will mimic this oracle procedure.

Algorithm 2 [Oracle Procedure 2]

1: Set $\mathcal{U} = \emptyset$.
2: Check whether $d_1^2/d_2^2 < 1 + c_n$, where $c_n > 0$ is some threshold depending on n (to be specified). If yes, add both $\mathbf{u}_1$ and $\mathbf{u}_2$ to $\mathcal{U}$ and go to Step 4; If no, go to Step 3.
3: Check whether $\mathbf{u}_1$ has two distinct values in its components. If yes, add $\mathbf{u}_1$ to $\mathcal{U}$ and go to Step 4; If no, add $\mathbf{u}_2$ to $\mathcal{U}$ and go to Step 4.
4: Return $\mathcal{U}$.
5: Use eigenvector(s) in $\mathcal{U}$ for clustering.

In step 2 of Oracle Procedure 2, a positive sequence c_n is to help check whether d_1^2 and d_2^2 are close enough. We include a buffer c_n because, in implementation, d_1 and d_2 are estimated with errors. As discussed above, the rationale behind step 3 is that when the second eigenvalue is much smaller than the first one, and so the estimated second eigenvector can be too noisy to be included for clustering, we use the estimated second eigenvector only when the first one is not usable. Oracle Procedure 2 prepares us to introduce our final practical selection procedure.

3.2 Comparison with a centering procedure

We digress here to discuss an existing procedure that drops an eigenvector. Concretely, a few works, such as IF-PCA, employ a step to first subtract the mean from the data. As will be demonstrated next, this approach reduces the second largest eigenvalue to 0 under our model, and thus always only uses the leading eigenvector for clustering. This can be advantageous under special conditions. However, we will also provide examples where our approach is superior. For this reason, we choose not to consider the centering procedure in detail in our paper.

Let $\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i$ and recall Model 1. By subtracting the expectation $\mathbb{E}\bar{\mathbf{x}} = \frac{n_1\boldsymbol{\mu}_1}{n} + \frac{n_2\boldsymbol{\mu}_2}{n}$, the model becomes

$$\mathbf{x}_i - \mathbb{E}\bar{\mathbf{x}} = Y_i\frac{n_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)}{n} + (1 - Y_i)\frac{n_1(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)}{n} + \mathbf{w}_i,\ i = 1, \ldots, n\,, \tag{5}$$

from which we can derive that

$$\mathrm{rank}(\mathbf{C}) := \mathrm{rank}\left((\mathbb{E}\mathbf{X} - (\mathbb{E}\bar{\mathbf{x}})\mathbf{1}_n^\top)(\mathbb{E}\mathbf{X} - (\mathbb{E}\bar{\mathbf{x}})\mathbf{1}_n^\top)^\top\right) = \mathrm{rank}\left(\frac{n_1n_2}{n}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\right) = 1\,.$$

Hence, the second eigenvalue of $\mathbf{C}$ is always 0, and the first eigenvalue, denoted by $d_1^2(\mathbf{C})$, is

$$d_1^2(\mathbf{C}) = \frac{n_1n_2\|\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2\|_2^2}{n}\,. \tag{6}$$

Comparing (6) with (3) and (4), we see that the effect of the demean step can be complicated. For one example, if $\boldsymbol{\mu}_1 = -\boldsymbol{\mu}_2$ and $n_1 = n_2$, then $d_1^2 = nc_{11}$, $\mathbb{E}\bar{\mathbf{x}} = 0$ and $d_1^2(\mathbf{C}) = nc_{11} = d_1^2$. In this case the demean approach is appropriate. On the other hand, if $c_{12} = 0$, then $d_1^2 = \max\{n_1c_{11}, n_2c_{22}\}$ and $d_2^2 = \min\{n_1c_{11}, n_2c_{22}\}$, whereas (6) becomes $d_1^2(\mathbf{C}) = \frac{n_1n_2(c_{11}+c_{22})}{n}$, which

lies between d_1^2 and d_2^2. Therefore in this case, the demean approach shrinks the first eigenvalue, reducing the signal strength (cf. the discussion after Oracle Procedure 1).

3.3 Eigen Selection Algorithm

The two oracle procedures discussed in Subsection 3.1 assume the knowledge of $\mathbf{H}$. In practice, we observe $\mathbf{X}$ instead of $\mathbf{H}$. Next, we will elevate our reasoning on $\mathbf{H}$ to that on $\mathbf{X}$ and propose an implementable algorithm for eigenvector selection. Denote by $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$ the eigenvectors of the matrix

$$\widehat{\mathbf{H}} := \mathbf{X}^\top \mathbf{X}\,,$$

corresponding to the two largest eigenvalues $\widehat{t}_1^2$ and $\widehat{t}_2^2$ ($\widehat{t}_1 \geq \widehat{t}_2 \geq 0$) of $\widehat{\mathbf{H}}$, respectively. As discussed after Oracle Procedure 1, $\widehat{t}_1$ and $\widehat{t}_2$ are the top singular values of $\mathbf{X}$, and d_1 and d_2 are the top singular values of $\mathbb{E}\mathbf{X}$. Thus, $\widehat{t}_1^2$ and $\widehat{t}_2^2$ estimate d_1^2 and d_2^2, respectively. Also, recall that $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$ are the top two right singular vectors of $\mathbf{X}$, while $\mathbf{u}_1$ and $\mathbf{u}_2$ are the top two right singular vectors of $\mathbb{E}\mathbf{X}$. Under some conditions, when $d_1^2/d_2^2 \neq 1$, i.e., no multiplicity, we have $\widehat{\mathbf{u}}_1(i) \approx \mathbf{u}_1(i)$ and $\widehat{\mathbf{u}}_2(i) \approx \mathbf{u}_2(i)$. Moreover, when $d_1^2 = d_2^2$, it is only possible for us to show that $(\widehat{\mathbf{u}}_1, \widehat{\mathbf{u}}_2) \approx (\mathbf{u}_1, \mathbf{u}_2)\widehat{\mathbf{U}}$ (e.g., by Davis-Kahan Theorem), where $\widehat{\mathbf{U}}$ is some 2×2 orthogonal matrix. Spectral clustering clusters $\mathbf{x}_i$'s into two groups by dividing the coordinates of $\widehat{\mathbf{u}}_1$ (and\or $\widehat{\mathbf{u}}_2$) into two groups via the k-means algorithm. In some scenarios, d_2 is small (compared to d_1) and $\widehat{\mathbf{u}}_2$ is significantly disturbed by the noise matrix $\mathbf{X} - \mathbb{E}\mathbf{X}$; in these scenarios, $\widehat{\mathbf{u}}_2$ is likely not good enough to distinguish the memberships. Putting these observations together, Oracle Procedure 2 can be implemented by replacing $(d_i, \mathbf{u}_i)$ with the sample version $(\widehat{t}_i, \widehat{\mathbf{u}}_i)$, $i = 1, 2$.

Based on the discussions above, we propose Algorithm 3: **E**igen-**S**elected **S**pectral **C**lustering Algorithm (ESSC). Let τ_n and δ_n be two diminishing positive sequences (i.e., $\tau_n + \delta_n = o(1)$) and $\mathbf{1}_n$ be an n-dimensional vector in which all entries are 1. In numerical implementation, we choose $\tau_n = \log^{-1}(n+p)$ and $\delta_n = \log^{-2}(n+p)$, which are guided by Theorems 2–3. Moreover, let

$$\mathfrak{f}_k = n^{-1/2}|\mathbf{1}_n^\top \widehat{\mathbf{u}}_k| - 1\,. \tag{7}$$

Note that if all entires of the unit vector $\mathbf{u}_1$ are equal, then $n^{-1/2}|\mathbf{1}_n^\top \mathbf{u}_1| = |\mathbf{u}_1^2(1) + \ldots + \mathbf{u}_1^2(n)| = 1$.

Hence, checking whether $|\mathfrak{f}_1|$ is small enough (e.g., $|\mathfrak{f}_1| < \delta_n$) is a reasonable substitute for checking whether $\mathbf{u}_1$ has all equal entries.

Algorithm 3 [Eigen-Selected Spectral Clustering (ESSC)]

1: Set $\widehat{\mathcal{U}} = \emptyset$.
2: Calculate $\widehat{t}_1$ and $\widehat{t}_2$ and the corresponding eigenvectors $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$ from $\widehat{\mathbf{H}}$.
3: Check whether $\widehat{t}_1/\widehat{t}_2 < 1 + \tau_n$. If yes, add both $\widehat{\mathbf{u}}_1$ and $\widehat{\mathbf{u}}_2$ to $\widehat{\mathcal{U}}$ and go to Step 5; if no, go to Step 4.
4: Check if $|\mathfrak{f}_1| \geq \delta_n$. If yes, add $\widehat{\mathbf{u}}_1$ to $\widehat{\mathcal{U}}$ and go to Step 5; if no, add $\widehat{\mathbf{u}}_2$ to $\widehat{\mathcal{U}}$ and go to Step 5.
5: Return $\widehat{\mathcal{U}}$.
6: Apply the k-means algorithm to vector(s) in $\widehat{\mathcal{U}}$ to cluster n instances into two groups.

4. THEORY OF ESSC

In this section, we derive a few theoretical results that support the steps 3 and 4 of Algorithm 3. We first prove in Proposition 1 asymptotic expansions for eigenvalues $\widehat{t}_1$ and $\widehat{t}_2$. In addition to motivating our handling of multiplicity as discussed in the previous section, these results potentially allow us to design a thresholding procedure on either $\widehat{t}_1 - \widehat{t}_2$ or $\widehat{t}_1/\widehat{t}_2$ to detect the multiplicity of eigenvalues. Indeed, our proposition fully characterizes the behavior of $\widehat{t}_1$ and $\widehat{t}_2$, so that we can derive an expansion for $\widehat{t}_1 - \widehat{t}_2$, but this expansion depends on the covariance matrix $\mathbf{\Sigma}$ (see Section S.4 in the Supplementary Material), which is not easy to estimate without the class label information. Similarly, an expansion of $\widehat{t}_1/\widehat{t}_2$ would involve $\mathbf{\Sigma}$. These concerns motivate us to resort to a less accurate but empirically feasible detection rule for eigenvalue multiplicity. Concretely, we derive concentration results regarding $\widehat{t}_1/\widehat{t}_2$, which do not rely on estimates of $\mathbf{\Sigma}$ and they give rise to step 3 of Algorithm 3. Theorems 2–3 provide a guarantee for using diminishing positive sequences τ_n and δ_n as thresholds for steps 3 and 4 in Algorithm 3. We adopt the following assumption in the theory section.

Assumption 1. *(i) The eigenvalues of* $\mathbf{\Sigma}$ *are bounded away from* 0 *and* ∞. *(ii)* $n^{1/C} \leq p \leq n^C$ *for some constant* $C > 0$.

Before presenting Proposition 1, we will introduce population quantities t_1 and t_2, which are asymptotically equivalent to population eigenvalues d_1 and d_2. We will establish below that t_1 and t_2 are indeed the asymptotic means of $\widehat{t}_1$ and $\widehat{t}_2$, respectively. As we work on $\mathcal{Z}$, a linearization of

$\widehat{\mathbf{H}}$, we will investigate $\mathbb{E}\mathcal{Z}$ and $\mathcal{Z}-\mathbb{E}\mathcal{Z}$. By Lemma 6 in the Supplementary Material, $\pm d_1$ and $\pm d_2$ are the eigenvalues of $\mathbb{E}\mathcal{Z}$, and the vector consisting of the first n entries of the eigenvector of $\mathbb{E}\mathcal{Z}$ corresponding to d_k equals $\frac{\mathbf{u}_k}{\sqrt{2}}$, $k=1,2$. Let the eigen decomposition of $\mathbb{E}\mathcal{Z}$ be

$$\mathbb{E}\mathcal{Z}=\Big[d_1(\mathbf{v}_1\mathbf{v}_1^\top-\mathbf{v}_{-1}\mathbf{v}_{-1}^\top)+d_2(\mathbf{v}_2\mathbf{v}_2^\top-\mathbf{v}_{-2}\mathbf{v}_{-2}^\top)\Big]\,,$$

in which $\mathbf{v}_1$ and $\mathbf{v}_2$ are the unit eigenvectors corresponding to d_1 and d_2, $\mathbf{v}_{-1}$ and $\mathbf{v}_{-2}$ are the unit eigenvectors corresponding to $-d_1$ and $-d_2$.

Define $\mathbf{V}=(\mathbf{v}_1,\mathbf{v}_2)$, $\mathbf{V}_-=(\mathbf{v}_{-1},\mathbf{v}_{-2})$ and $\mathbf{D}=\mathrm{diag}(d_1,d_2)$. Then the eigen decomposition of $\mathbb{E}\mathcal{Z}$ can be written as

$$\mathbb{E}\mathcal{Z}=\mathbf{V}\mathbf{D}\mathbf{V}^\top-\mathbf{V}_-\mathbf{D}\mathbf{V}_-^\top\,. \tag{8}$$

Moreover, let

$$\mathbf{W}=\mathcal{Z}-\mathbb{E}\mathcal{Z}=\begin{pmatrix}0 & (\mathbf{X}-\mathbb{E}\mathbf{X})^\top\\ \mathbf{X}-\mathbb{E}\mathbf{X} & 0\end{pmatrix}. \tag{9}$$

For complex variable z, and any matrices (or vectors) $\mathbf{M}_1$ and $\mathbf{M}_2$ of suitable dimensions, we define the following notations.

$$\mathcal{R}(\mathbf{M}_1,\mathbf{M}_2,z)=-\sum_{l=0,l\neq 1}^{L}z^{-(l+1)}\mathbf{M}_1^\top\mathbb{E}\mathbf{W}^l\mathbf{M}_2\,, \tag{10}$$

and

$$f(z)=\begin{pmatrix}f_{11}(z) & f_{12}(z)\\ f_{21}(z) & f_{22}(z)\end{pmatrix}=\mathbf{I}+\mathbf{D}\Big(\mathcal{R}(\mathbf{V},\mathbf{V},z)-\mathcal{R}(\mathbf{V},\mathbf{V}_-,z)\big(-\mathbf{D}+\mathcal{R}(\mathbf{V}_-,\mathbf{V}_-,z)\big)^{-1}\mathcal{R}(\mathbf{V}_-,\mathbf{V},z)\Big)\,. \tag{11}$$

Note that $f(z)$ is a crucial term to determine the locations of the eigenvalues $\widehat{t}_1$ and $\widehat{t}_2$. Indeed, $f(z)$ is the asymptotic non-random terms of the eigenvalue function (S.47) in the Supplementary Material.

Lemma 1. *Denote by $a_n=d_2-\sigma_n$ and $b_n=d_1+\sigma_n$. Assume that*

$$d_1-d_2=o(\sqrt{d_2})\ \text{and}\ d_2\gg\sigma_n^{4/3}\,, \tag{12}$$

then we have the following conclusions

1. *The equation*

$$\det(f(z)) = 0\,, \tag{13}$$

in which $f(z)$ is defined in (11), has at most two solutions in $[a_n, b_n]$. We denote these solutions by t_1 and t_2 with $t_2 \le t_1$.

2.

$$t_k - d_k = O\left(\frac{\sigma_n^2}{d_2}\right),\ k = 1, 2\,. \tag{14}$$

Equation (12) is a signal strength assumption requiring that the top two eigenvalues should be spiked enough, and that the second eigenvalue cannot be much smaller than the top eigenvalue. In fact, (12) implies that $d_1/d_2 \to 1$, that is, close to multiplicity. Under such conditions, Lemma 1 guarantees the existence of t_1 and t_2. Moreover, this lemma provides a guarantee that $\frac{t_1}{d_1}$ and $\frac{t_2}{d_2}$ are asymptotically close to 1. The following proposition is established by carefully analyzing the behavior of $\widehat{t}_k$ around t_k, $k = 1, 2$.

Proposition 1. *Under Assumption 1 and (12), we have*

$$\widehat{t}_1 - t_1 = \frac{1}{2}\left[-g_{11}(t_1) - g_{22}(t_1) + \left\{(g_{11}(t_1) + g_{22}(t_1))^2 - 4\left(g_{11}(t_1)g_{22}(t_1) - g_{12}^2(t_1)\right)\right\}^{\frac{1}{2}}\right] + o_p(1)\,, \tag{15}$$

$$\widehat{t}_2 - t_2 = \frac{1}{2}\left[-g_{11}(t_2) - g_{22}(t_2) - \left\{(g_{11}(t_2) + g_{22}(t_2))^2 - 4\left(g_{11}(t_2)g_{22}(t_2) - g_{12}^2(t_2)\right)\right\}^{\frac{1}{2}}\right] + o_p(1)\,, \tag{16}$$

where g_{11}, g_{12}, g_{21} and g_{22} are defined in

$$g(z) = \begin{pmatrix} g_{11}(z) & g_{12}(z) \\ g_{21}(z) & g_{22}(z) \end{pmatrix} = z^2\mathbf{D}^{-1}f(z) - \mathbf{V}^\top\mathbf{W}\mathbf{V}\,. \tag{17}$$

For $\widehat{t}_2$, we also have an alternative expression

$$\widehat{t}_2 - t_1 = \frac{1}{2}\left[-g_{11}(t_1) - g_{22}(t_1) - \left\{(g_{11}(t_1) + g_{22}(t_1))^2 - 4\left(g_{11}(t_1)g_{22}(t_1) - g_{12}^2(t_1)\right)\right\}^{\frac{1}{2}}\right] + o_p(1)\,. \tag{18}$$

By the arguments before Lemma 1 and (S.54), $g(z)$ is the matrix to determine $\widehat{t}_k - t_1$, $k = 1, 2$. Concretely, $\widehat{t}_k - t_1$ can be approximated by the eigenvalues of $g(z)$. Proposition 1 provides asymptotic expansions of $\widehat{t}_k$ around t_k $(k = 1, 2)$ that are not achievable by routine application of the Weyl's inequality. Indeed, Proposition 1 implies that the fluctuations of $\widehat{t}_k$ around t_k is $O_p(1)$ (cf., Lemma 2 in the Supplementary Material), while the Weyl's inequality gives $|\widehat{t}_k - d_k| \le \|\mathbf{W}\|$, which, combined with Lemma 4 in the Supplementary Material, implies that the fluctuation of $\widehat{t}_1 - \widehat{t}_2$ around $d_1 - d_2$ is $O_p(\sigma_n)$. On the other hand, Proposition 1 also suggests that designing a statistical procedure by thresholding $\widehat{t}_1 - \widehat{t}_2$ would be a difficult task, as explained in detail in Section S.4 of the Supplementary Material.

Similar to the asymptotic expansion for $\widehat{t}_1 - \widehat{t}_2$, an asymptotic expansion for $\widehat{t}_1/\widehat{t}_2$ would also involve the covariance matrix $\mathbf{\Sigma}$. Nevertheless, the latter has better concentration property compared to the former, which motivates us to consider a non-random thresholding rule on $\widehat{t}_1/\widehat{t}_2$. The concentration properties of $\widehat{t}_1/\widehat{t}_2$ under different population scenarios are summarized in Theorem 2 and the first part of Theorem 3, respectively, with the former corresponding to the case close to multiplicity and the latter corresponding to the case away from multiplicity. Moreover, the second part of Theorem 3 validates the step 4 of ESSC. We would like to emphasize that Theorem 3 does not require d_2 to be spiked and thus can be applied even when $d_2 = 0$.

Theorem 2. *In addition to Assumption 1, further assume that $d_1 \gg \sigma_n$, $d_1/d_2 \le 1 + n^{-c}$ for all $n \ge n_0$, where c and n_0 are positive constants, then there exists a positive constant C such that as $n \to \infty$,*

$$\mathbb{P}\left(\frac{\widehat{t}_1}{\widehat{t}_2} \ge 1 + C\left(\frac{1}{n^c} + \frac{\sigma_n}{d_1}\right)\right) \to 0\,. \tag{19}$$

Theorem 3. *In addition to Assumption 1, further assume that $d_1 \gg \sigma_n$ and $d_1/d_2 \ge 1 + c$ for*

some positive constant c. Then for any positive constant D, we have

$$\mathbb{P}\left(\frac{\widehat{t}_1}{\widehat{t}_2} \geq 1 + \frac{c}{2}\right) \geq 1 - n^{-D}, \tag{20}$$

for all $n \geq n_0$, where n_0 is some constant that only depends on the constant D. Moreover, if all the entries of $\mathbf{u}_1$ are equal, we have for all $n \geq n_0$,

$$\mathbb{P}\left(\left|\left(\frac{1}{n}\right)^{\frac{1}{2}} |\mathbf{1}_n^\top \widehat{\mathbf{u}}_1| - 1\right| \leq \sqrt{\frac{2\sigma_n}{d_1}}\right) \geq 1 - n^{-D}. \tag{21}$$

We note that Theorems 2 and 3 require $d_1 \gg \sigma_n$, which is weaker than the condition for d_1 in Proposition 1. Indeed, the weaker conditions in Theorems 2 and 3 are the sufficient conditions for our eigen selection procedure to work. By Theorems 2 and 3, we can choose τ_n and δ_n for Algorithm 3 such that $C(n^{-c} + \sigma_n/d_1) \leq \tau_n \leq c/2$ and $\delta_n \geq \sqrt{2\sigma_n/d_1}$. In our simulation, we let $\tau_n = \log^{-1}(n+p)$ and $\delta_n = \log^{-2}(n+p)$. These choices were reasonable when $\log^{-4}(n+p) \geq 2\sigma_n/d_1$ for sufficiently large n and p.

We next discuss that when $p \sim n$, the results in Theorems 2–3 apply as long as clustering is possible. Concretely, note that both these theorems require that d_1, a measure of the difficulty in clustering, to satisfy $d_1 \gg \sigma_n$, which reduces to $d_1 \gg \sqrt{n}$ when $p \sim n$. In the Supplementary Material, we establish the clustering lower bound in Theorem 5 by showing that if $d_1 \ll \sqrt{n}$, then clustering is impossible regardless of what method to use; see the Supplementary Material for specific assumptions. We further prove in Theorem 6 and Corollary 1 in the Supplementary Material that when $d_1 \geq \sqrt{2(1+\epsilon_0)n \log n}$ for any positive constant ϵ_0, a simple clustering method based on the signs of selected eigenvector can perfectly recover the class labels with probability tending to 1 (i.e., exact recovery). Our exact recovery result is similar to Theorem 3.1 of Abbe et al. [2020+], which studied symmetric random matrices with independent entries on and above diagonals and low expected rank. Moreover, in related papers working on different models such as $\mathbb{Z}_2$-synchronization [Bandeira et al., 2017] and stochastic block model [Abbe et al., 2020+], it is shown that when $d_1(\mathbf{A})$ is at least of order $\sqrt{n \log n}$, there exists an exact recovery approach to identify the memberships, where $\mathbf{A}$ is the data matrix in the respective context.

5. GESSC: EXTENSION TO MULTIPLE CLUSTERS

In this section, we consider K clusters with $K > 2$ and given. Suppose $\mathbf{x}$ follows a Gaussian mixture type model that has K different populations means:

$$\mathbf{x}_i = \sum_{j=1}^{K} \mathbb{1}(Y_i = j)\boldsymbol{\mu}_j + \mathbf{w}_i\,,\ i = 1, \ldots, n\,,$$

where $\mathbb{1}(\cdot)$ is the indicator function, $\mathbf{w}_i$ follows the assumptions in (1), and $\{Y_i\}_{i=1}^n$ are deterministic latent class labels taking values in $\{1, 2, \ldots, K\}$. Similar to the $K = 2$ scenario, we define n-dimensional latent vectors $\mathbf{a}_k$, $k = 1, 2, \ldots, K$, whose components are either 1 or 0. Concretely,

$$\mathbf{a}_k(i) = 1 \text{ if and only if } \mathbf{x}_i \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}),\ k = 1, 2, \ldots, K\,.$$

Moreover, we denote $n_k = \|\mathbf{a}_k\|_2^2$ and $c_{kl} = \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_l$, $1 \le k, l \le K$. Similar to the definition of $\mathbf{H}$ in (2), we define

$$\mathbf{H} \coloneqq (\mathbb{E}\mathbf{X})^\top \mathbb{E}\mathbf{X} = \sum_{1 \le k,l \le K} \mathbf{a}_k \mathbf{a}_l^\top c_{kl} \ge 0\,. \tag{22}$$

Note that $\operatorname{rank}(\mathbf{H})$ is not necessarily equal to K because the cluster centers $\boldsymbol{\mu}_k$, $k = 1, \cdots, K$ may be linearly dependent. Throughout this section, we denote $\operatorname{rank}(\mathbf{H}) = K_0 \le K$.

Let $\mathbf{U} = (\mathbf{u}_1, \ldots, \mathbf{u}_{K_0}, -\mathbf{u}_1, \ldots, -\mathbf{u}_{K_0})$, where $\mathbf{u}_k$ is the unit right singular vector of $\mathbb{E}\mathbf{X}$ corresponding to $d_k = \sigma_k(\mathbb{E}\mathbf{X})$. Further, let $\widehat{\mathbf{u}}_k$ be the unit right singular vector of $\mathbf{X}$ corresponding to $\sigma_k(\mathbf{X})$. We have the following theorem.

Theorem 4. *Under the condition that $d_{K_0} \gg \sigma_n$ and Assumption 1, for any positive constant D and n-dimensional unit vector $\mathbf{x}$, there exists K_0 orthogonal $2K_0$-dimensional unit vectors $(\mathbf{o}_1, \ldots, \mathbf{o}_{K_0})$ such that*

$$\mathbb{P}\left(\max_{1 \le k \le K_0} |\mathbf{x}^\top(\widehat{\mathbf{u}}_k - \mathbf{U}\mathbf{o}_k)| \le \sqrt{\frac{2\sigma_n}{d_{K_0}}}\right) \ge 1 - n^{-D}. \tag{23}$$

Moreover, for any $1 \le k \le K_0$, we have $\mathbf{o}_k = \mathbf{e}_k$ or $-\mathbf{e}_k$ if $\mathbf{u}_k$ has multiplicity 1 and $\min\{d_k/d_{k+1}, d_{k-1}/d_k\}$ is bounded away from 1 by some positive constant c, where $d_0 \coloneqq \infty$.

Take $\mathfrak{f}_k = n^{-1/2}|\mathbf{1}_n^\top \widehat{\mathbf{u}}_k| - 1$ in (7). A small $\mathfrak{f}_k$ means $|\left(\frac{1}{n}\right)^{1/2} \mathbf{1}_n^\top \widehat{\mathbf{u}}_k|$ is close to 1. Combining

this with Theorem 4, a small $\mathfrak{f}_k$ implies the closeness between $\left(\frac{1}{n}\right)^{1/2}|\mathbf{1}_n^\top \mathbf{U}\mathbf{o}_k|$ and 1 with high probability. This together with Cauchy-Schwarz inequality further implies that the entries of $\mathbf{U}\mathbf{o}_k$ are close to each other with high probability, and thus $\widehat{\mathbf{u}}_k$ should be screened out because of no/low clustering power. This motivates the following algorithm.

Algorithm 4 [Generalized Eigen-Selected Spectral Clustering (GESSC)]

Input: K_0.

1: Set $\widehat{\mathcal{U}} = \emptyset$.
2: For $1 \le k \le K_0$, check if $|\mathfrak{f}_k| \ge \delta_n$. If yes, add $\widehat{\mathbf{u}}_k$ to $\widehat{\mathcal{U}}$.
3: Return $\widehat{\mathcal{U}}$.
4: Apply the k-means algorithm to vector(s) in $\widehat{\mathcal{U}}$ to cluster n instances into K groups.

Algorithm 4 is not immediately implementable because even though K is given, K_0 is unknown yet. Motivated by the rank estimation approach in Fan et al. [2020], we propose a similar algorithm to estimate K_0. Concretely, let $\Phi = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\mathbf{x}_i \mathbb{E}\mathbf{x}_i^\top$ and $\widehat{\Phi} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$. Further define $\widehat{\mathbf{R}} = \mathrm{diag}(\widehat{\Phi})^{-1/2}\,\widehat{\Phi}\,\mathrm{diag}(\widehat{\Phi})^{-1/2}$ and its corresponding population version $\mathbf{R} = \mathrm{diag}(\Phi)^{-1/2}\,\Phi\,\mathrm{diag}(\Phi)^{-1/2}$. Compared with Fan et al. [2020], we skipped the demean step in defining Φ, $\widehat{\Phi}$, $\mathbf{R}$ and $\widehat{\mathbf{R}}$. The intuition of this rank estimation approach is that the rank of Φ and $\mathbf{R}$ are both equal to K_0, which motivates us to estimate K_0 by examining spiked eigenvalues of $\widehat{\Phi}$ or $\widehat{\mathbf{R}}$. However, the magnitude of spiked eigenvalues of $\widehat{\Phi}$ is not scale free and depends on the unknown population parameters $\mathbf{\Sigma}$. Thanks to the scaling step in $\mathbf{R}$ and $\widehat{\mathbf{R}}$, the top K_0 spiked eigenvalues can be separated from the remaining ones by some threshold independent of the unknown population parameter. The following algorithm is based on this intuition.

Algorithm 5 [Estimation of K_0]

1: Let $\lambda_j^{\mathbf{C}}(\widehat{\mathbf{R}}) = -1/\underline{m}_{n,j}(\lambda_j(\widehat{\mathbf{R}}))$, where $\underline{m}_{n,j}(z) = -(1-(p-j)/n)z^{-1} + (p-j)/n m_{n,j}(z)$ and $m_{n,j}(z) = \left(\sum_{i=j+1}^p (\lambda_i(\widehat{\mathbf{R}}) - z)^{-1} + ((3\lambda_j(\widehat{\mathbf{R}}) + \lambda_{j+1}(\widehat{\mathbf{R}}))/4 - z)^{-1}\right)/(p-j)$.
2: $\widehat{K}_0 = \max\{j \in \{1,\dots,K\} :\ \lambda_j^{\mathbf{C}}(\widehat{\mathbf{R}}) > 1 + \sqrt{p/n}\}$.

The term $\lambda_j^{\mathbf{C}}(\widehat{\mathbf{R}})$ in Algorithm 5 is the bias-corrected version of the jth eigenvalue of $\widehat{\mathbf{R}}$, and the threshold that can separate the top K_0 spiked eigenvalues from the rest is $1 + \sqrt{p/n}$, which is the same as the one proposed in Fan et al. [2020]. The consistency of $\widehat{K}_0$ is ensured by Proposition 5 below.

Assumption 2. $\|\mathrm{diag}(\Phi)^{-1/2}\,\Sigma\| \le 1$, $\lambda_{K_0}(\mathbf{R}) \to \infty$, $p/n \to c$ *for some positive constant* c.

Assumption 3. *The diagonal entries of* Φ *are bounded away from* 0 *and infinity. Moreover, the limiting spectral distribution of* $\mathrm{diag}(\Phi)^{-1/2}\,\mathbf{\Sigma}\,\mathrm{diag}(\Phi)^{-1/2}$ *exists and we denote it by* $H(t)$.

Theorem 5. *Let* $\widehat{K}_0$ *be returned from Algorithm 5. Under Assumptions 1(i), 2–3, we have*

$$\mathbb{P}(\widehat{K}_0 = K_0) = 1 - o(1)\,.$$

Moreover, we have

$$\mathbb{P}\left(\max_{1\le k\le \widehat{K}_0} |\mathbf{x}^\top(\widehat{\mathbf{u}}_k - \mathbf{U}\mathbf{o}_k)| \le \sqrt{\frac{2\sigma_n}{d_{\widehat{K}_0}}}\right) = 1 - o(1)\,, \tag{24}$$

where $\mathbf{o}_k$ *is the same as Theorem 4.*

Now we compare Algorithm 4 ($K > 2$) and Algorithm 3 ($K = 2$). The major distinction is that Algorithm 3 screens out two types of eigenvectors: 1) the one without clustering power and 2) the one that cannot be estimated accurately, while Algorithm 4 only screens out the first type. This is because when $K = 2$, a useful eigenvector has full clustering power. But when $K > 2$, a useful eigenvector may only have partial clustering power in the sense that some clusters may collapse along that eigenvector direction. Thus, despite that some eigenvectors may not be able to be estimated accurately, including them can still be beneficial because they may carry important information on certain clusters that are not provided by other eigenvectors.

6. SIMULATION STUDIES

In this section, we first compare our newly proposed eigen-selected spectral clustering (ESSC) with k-means, Spectral Clustering, CHIME, IF-PCA and the oracle classifier (a.k.a, Bayes classifier) under Models 1-5. Here k-means algorithm means that we directly apply the k-means to the data $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. Recall that the oracle classifier to distinguish $\mathbf{x}|(Y = 1) \sim N(\boldsymbol{\mu}_1, \mathbf{\Sigma})$ from $\mathbf{x}|(Y = 0) \sim N(\boldsymbol{\mu}_2, \mathbf{\Sigma})$ is

$$g(\mathbf{x}) = \begin{cases} 1, & \text{if } (\mathbf{x} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \ge \log(\frac{\pi}{1-\pi})\,, \\ 0, & \text{if } (\mathbf{x} - \frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_2}{2})^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) < \log(\frac{\pi}{1-\pi})\,, \end{cases} \tag{25}$$

where $\pi = \mathbb{P}(Y = 1)$. We generate n i.i.d. copies of $\mathbf{x} \sim \pi N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}) + (1-\pi)N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ with $\pi = 0.5$. We have also experimented with $\pi = 0.4$ and the results are very similar so omitted. Throughout this section, we set $\boldsymbol{\mu}_1 = r(\boldsymbol{\mu}_{11}^\top, \boldsymbol{\mu}_{12}^\top)^\top$, where $\boldsymbol{\mu}_{11}$ is an l-dimensional vector in which all entries are 1, $\boldsymbol{\mu}_{12}$ is a $(p-l)$-dimensional vector in which all entries are 0, and r is a scaling parameter. Then under Model 6, we compare GESSC with other methods. The simulation models are specified as follows.

- Model 1: $\boldsymbol{\mu}_2 = \mathbf{0}$, $n = 200$, $p \in \{100, 200, 400, 600, 800, 1000, 1200\}$, $l = 15$ and $r = 2$. The covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij})$ is symmetric with $\boldsymbol{\Sigma}_{ij} = 0.8^{|i-j|}$.
- Model 2: $\boldsymbol{\mu}_2 = r(\boldsymbol{\mu}_{12}^\top, \boldsymbol{\mu}_{11}^\top)^\top$, $n = 100$, $p \in \{100, 200, 400, 600, 800, 1000, 1200\}$, $l = 12$ and $r = 2$. The covariance matrix $\boldsymbol{\Sigma} = r^2\mathbf{I}$.
- Model 3: $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1/2$, $n = 200$, $p \in \{100, 200, 400, 600, 800, 1000, 1200\}$, $l = 60$ and $r = 1$. The covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}$.
- Model 4: the same as Model 3 except for $p \in \{30, 50, 100, 200, 400, 600, 800\}$ and $l = 30$.
- Model 5: $\boldsymbol{\mu}_2 = 1/r(\boldsymbol{\mu}_{21}^\top, \boldsymbol{\mu}_{22}^\top)^\top$, where $\boldsymbol{\mu}_{21}$ is an $(l/2)$-dimensional vector in which all entries are 1, $\boldsymbol{\mu}_{22}$ is a $(p - l/2)$-dimensional vector in which all entries are 0, $l = 20$, $p = 400$, $n \in \{200, 400, 600, 800, 1000\}$ and $r = 1$. The covariance matrix $\boldsymbol{\Sigma} = r^2\mathbf{I}$.
- Model 6: $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_1/2$, $\boldsymbol{\mu}_3 = 0$, $n = 100$, $p \in \{100, 200, 400, 600, 800, 1000, 1200\}$, $l = 20$ and $r = 2$. The covariance matrix $\boldsymbol{\Sigma} = r\mathbf{I}$.

In Model 1, the covariance matrix $\boldsymbol{\Sigma}$ has non-zero off-diagonal entries. In Models 2–4, each non-zero entry of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ with magnitude not larger than r is covered by Gaussian noise with variance r^2. In Models 3–4, $\boldsymbol{\mu}_1$ is parallel to $\boldsymbol{\mu}_2$. With Model 5, we investigate how the trend of the misclustering rate changes with n. Model 6 is the case of multiple clusters.

For CHIME , we use the Matlab codes uploaded to `Github` by the authors of Cai et al. [2013]. Since CHIME involves an EM algorithm, the initial value is very important. We use the default initial values provided in the Matlab codes. We also need to provide the other initial values of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\beta_0 = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and π denoted by $\widehat{\boldsymbol{\mu}}_1$, $\widehat{\boldsymbol{\mu}}_2$, $\widehat{\beta}_0$ and $\widehat{\pi}$ respectively. Specifically, we

set $\widehat{\boldsymbol{\mu}}_1 = \frac{\sum_{1\le i\le n, Y_i=1}\mathbf{x}_i}{n_1}$ and $\widehat{\boldsymbol{\mu}}_2 = \frac{\sum_{1\le i\le n, Y_i=0}\mathbf{x}_i}{n_2}$, $\widehat{\beta}_0 = \boldsymbol{\Sigma}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_2)$ and $\widehat{\pi} = 0.4$. For Spectral Clustering, there are a lot of variants. In the simulation part, we follow Ng et al. [2002] with the common non-linear kernel $k(\mathbf{x}, \mathbf{y}) = \exp\{-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{2p}\}$ to construct an affinity matrix. For IF-PCA in Jin and Wang [2016], we directly apply the Matlab code provided by the authors without modification.

We repeat 100 times for each model setting and calculate the average misclustering rate and the corresponding standard error in Tables 3-8.

Table 3: The misclustering rate of several approaches for Model 1 with $\pi = 0.5$

p	ESSC	k-means	Spectral Clustering	CHIME	IF-PCA	Oracle
100	.067(.0017)	.069(.0018)	.071(.0017)	.036(.0045)	.14(.0112)	.002(.0009)
200	.072(.0017)	.074(.0019)	.076(.0019)	.071(.0097)	.15(.0131)	.002(.001)
400	.073(.0021)	.079(.0022)	.081(.0021)	.088(.0125)	.191(.0137)	.002(.0009)
600	.078(.002)	.088(.0022)	.091(.0022)	.067(.0105)	.21(.0146)	.002(.001)
800	.078(.0018)	.1(.0055)	.099(.0023)	.036(.0047)	.258(.0157)	.002(.001)
1000	.084(.002)	.117(.0063)	.108(.0026)	.024(.0046)	.257(.0149)	.002(.0009)
1200	.087(.0022)	.12(.0053)	.117(.003)	.021(.005)	.266(.0147)	.002(.0009)

Table 4: The misclustering rate of several approaches for Model 2 with $\pi = 0.5$

p	ESSC	k-means	Spectral Clustering	CHIME	IF-PCA	Oracle
100	.012(.0011)	.011(.001)	.083(.013)	.004(.0006)	.224(.0139)	.008(.0008)
200	.023(.0016)	.024(.004)	.169(.015)	.002(.0004)	.269(.0139)	.007(.0008)
400	.042(.0029)	.04(.0049)	.298(.013)	0(0)	.335(.0124)	.009(.0009)
600	.068(.0034)	.089(.0103)	.352(.0096)	0(0)	.373(.0107)	.007(.0007)
800	.086(.0037)	.122(.0121)	.386(.0073)	0(0)	.401(.0088)	.006(.0007)
1000	.117(.0057)	.211(.0145)	.386(.0078)	0(0)	.423(.0076)	.008(.001)
1200	.16(.0084)	.238(.0142)	.398(.0069)	0(0)	.407 (.0071)	.006(.0009)

Table 5: The misclustering rate of several approaches for Model 3 with $\pi = 0.5$

p	ESSC	k-means	Spectral Clustering	CHIME	IF-PCA	Oracle
100	.028(.0012)	.037(.0014)	.038(.0014)	.093(.0121)	.203(.0096)	.028(.0012)
200	.028(.0011)	.047(.0014)	.049(.0013)	.438(.0117)	.285(.0117)	.026(.0012)
400	.027(.001)	.085(.0075)	.073(.0023)	.446(.0106)	.366(.0107)	.026(.001)
600	.032(.0014)	.137(.011)	.1(.0023)	.468(.0049)	.393(.0088)	.025(.0012)
800	.033(.0013)	.193(.011)	.134(.0034)	.442(.0109)	.41(.008)	.029(.0012)
1000	.033(.0015)	.269(.0127)	.161(.004)	.457(.0082)	.424(.0066)	.026(.0012)
1200	.037(.0013)	.322(.0114)	.196(.0059)	.365(.0118)	.425(.0071)	.026(.0011)

In general, ESSC deteriorates much slower than k-means as p increases and is more stable than

k-means. Tables 3–4 indicate that k-means is comparable to ESSC when p is small, while ESSC works better than k-means when p is large. For Model 3 in Table 5, ESSC outperforms k-means. Since the number of non-zero coordinates of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ in Model 4 is much fewer than that in Model 3, the signal strength of the means in Model 4 is not strong enough to have large spiked singular values. As such, the performance of ESSC in Table 6 is worse than that of k-means when p is smaller (e.g., less than 200). However, since the misclustering rate of ESSC increases slowly as p increases, when p passes 200, ESSC competes favorably against k-means. Comparing to Spectral Clustering, ESSC excels in all models for almost all p and n. Tables 3–4 indicate that CHIME outperforms the other approaches for Models 1–2. While for Models 3–4, the performance of CHIME is worse than the others. We conjecture that such a phenomenon happens because the differences of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are small and $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ has more non-zero coordinates than that in Model 2, which does not cater the sparse assumptions in CHIME very well. Table 7 for Model 5 indicates how the misclustering rates change as n increases. When n is small, We also observe that ESSC performs better than other methods. Since CHIME is designed for the case of two clusters and we do not run CHIME in Table 8. Table 8 shows that GESSC outperforms the other approaches for Model 6.

Table 6: The misclustering rate of several approaches for Model 4 with $\pi = 0.5$

p	ESSC	k-means	Spectral Clustering	CHIME	IF-PCA	Oracle
30	.19(.003)	.105(.0023)	.103(.002)	.47(.0024)	.235(.0055)	.087(.0021)
50	.2(.0033)	.112(.003)	.111(.0026)	.472(.0021)	.301(.0083)	.088(.0019)
100	.21(.003)	.145(.0059)	.133(.0029)	.474(.002)	.341(.009)	.084(.0018)
200	.21(.0028)	.24(.0107)	.182(.0048)	.474(.0022)	.419(.0065)	.086(.0018)
400	.23(.0031)	.372(.008)	.279(.0079)	.471(.0019)	.448(.0041)	.086(.0019)
600	.241(.0034)	.41(.006)	.348(.0075)	.47(.0023)	.452(.004)	.086(.002)
800	.255(.0034)	.419(.0059)	.349(.0071)	.473(.0021)	.46(.0026)	.088(.002)

Table 7: The misclustering rate of several approaches for Model 5 with $\pi = 0.5$

n	ESSC	k-means	Spectral Clustering	CHIME	IF-PCA	Oracle
200	.04(.0015)	.073(.0058)	.347(.0096)	.079(.0007)	.384(.0108)	.014(.0009)
400	.033(.0009)	.042(.0012)	.191(.0137)	.016(.0006)	.305(.0133)	.015(.0006)
600	.03(.0007)	.036(.0008)	.062(.0067)	.022(.0007)	.288(.0139)	.013(.0004)
800	.029(.0007)	.032(.0007)	.037(.0021)	.029(.0006)	.291(.0147)	.013(.0004)
1000	.029(.0005)	.031(.0005)	.033(.0008)	.034(.0006)	.28(.0154)	.014(.0004)

Table 8: The misclustering rate of several approaches for Model 6

p	GESSC	k-means	Spectral Clustering	IF-PCA
100	.099(.0029)	.239(.0076)	.326(.0064)	.481(.0075)
200	.108(.0035)	.309(.0074)	.343(.0053)	.527(.0076)
400	.12(.0047)	.339(.0063)	.356(.0053)	.547(.0073)
600	.138(.0061)	.363(.0049)	.357(.0044)	.559(.0063)
800	.18(.0088)	.399(.0065)	.378(.0052)	.567(.0061)
1000	.2(.0088)	.416(.0065)	.396(.0056)	.575(.0058)
1200	.255(.0091)	.436(.0067)	.399(.0048)	.584(.0049)

7. REAL DATA ANALYSIS

In this section, we run several real data sets in finance and biomedical diagnosis to compare the newly proposed ESSC with the other clustering approaches.

7.1 Financial data

We consider a credit card dataset in ULB and Worldline [2018]. This dataset contains transactions made by credit cards in September 2013 by European cardholders. Each instance in the data contains 30 features and the data has labeled 492 frauds out of $284,807$ transactions. Among these features, 28 are engineered features obtained from some original features (which are not revealed for privacy concerns), while the other two features are 'Time' and 'Amount'. We only use the 28 engineered features to do clustering. Clearly, the data set is highly imbalanced: the fraud transactions account for 0.172% of all transactions. We choose the first 50 fraud transactions and the first $5r$ normal transactions, where $r \in \{10, 11, \ldots, 50\}$. Note that for $r = 10$, the fraud and normal groups are balanced in size, and for $r = 50$, normal transactions are 5 times as many as the fraud ones. On these data sets, we compare ESSC with IF-PCA and two other spectral clustering methods. The first spectral method (SC1) directly applies k-means to the first n rows of $(\widehat{\mathbf{v}}_1, \widehat{\mathbf{v}}_2)$ and the second method (SC2) is the one that uses a non-linear kernel as described in the simulation section. We do not report the performance of CHIME in real data analysis, as initializations on parameters such as Σ are not communicated in the original paper and unlike simulation, there is no obvious initialization choice for real data studies. Figure 1 demonstrates that ESSC is the preferred approach for all r's (i.e., imbalanced ratios), demonstrating the efficiency and stability of ESSC on this financial data set.

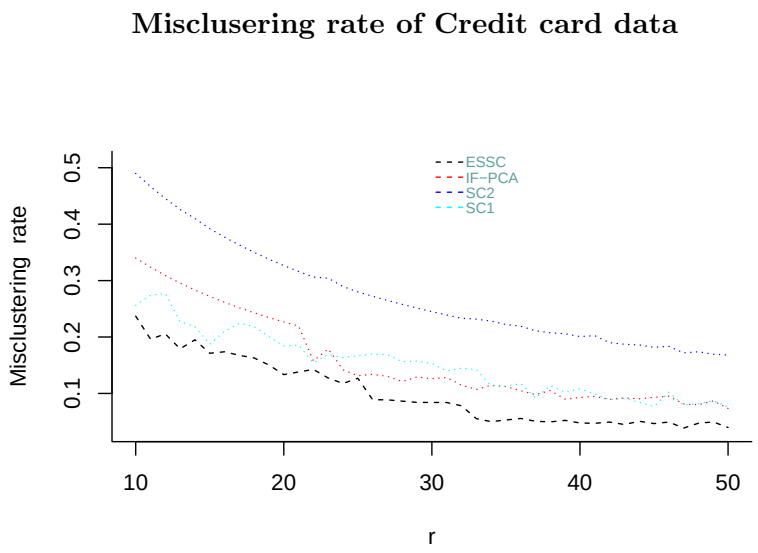

Figure 1: Misclustering rate of the Credit card data vs. different sample sizes $n = 5(r + 10)$. The red curve represents IF-PCA, the cyan curve represents SC1, the blue curve represents SC2 and the black curve represents ESSC.

7.2 Biological data

We use several gene microarray data sets collected and processed by authors in Jin and Wang [2016]. These data sets are canonical datasets analyzed in the literature such as in Dettling [2004], Gordon et al. [2002] and Yousefi et al. [2009]. We use a processed version at www.stat.cmu.edu/ jiashun/Research/software/GenomicsData. We apply the four approaches mentioned in the financial data section. All the datasets considered in this section belong to the ultra-high-dimensional settings. In each dataset, the number of features is about two orders of magnitude larger than the sample size; see Table 9 for a summary. In supervised learning, when feature dimensionality and sample size have such a relation, some independence screening procedure is usually beneficial before implementing methods from joint modeling. We will adopt a similar two-step pipeline for clustering. As IF-PCA involves an independence screening step via normalized KS-statistic ((1.7) of Jin and Wang [2016]), we also implement this screening step before calling other methods. Concretely on each dataset, for each $p \in \{150, 151, 152, \ldots, 300\}$, we keep the p features that have the largest p normalized KS-statistic and construct a $p \times n$ matrix $\mathbf{X}$. Then, since the dimension reduction step is done, for IF-PCA we only apply the "PCA-2" step in Jin and Wang [2016]. Moreover, we subsample each dataset so that the resulting datasets all have an average size of 60. Concretely,

when a dataset has n instances, we keep each instance with a probability $60/n$. For each dataset, we repeat the subsampling procedure 10 times and report the average misclustering rates of the clustering methods on the subsamples.

Table 9: Sample size and dimensionality of real data sets

Data Name	Sample size	Total number of features
Colon Cancer	62	2000
Breast Cancer	276	22215
Lung Cancer 1	203	12600
Lung Cancer 2	181	12533
Leukemia	72	3571

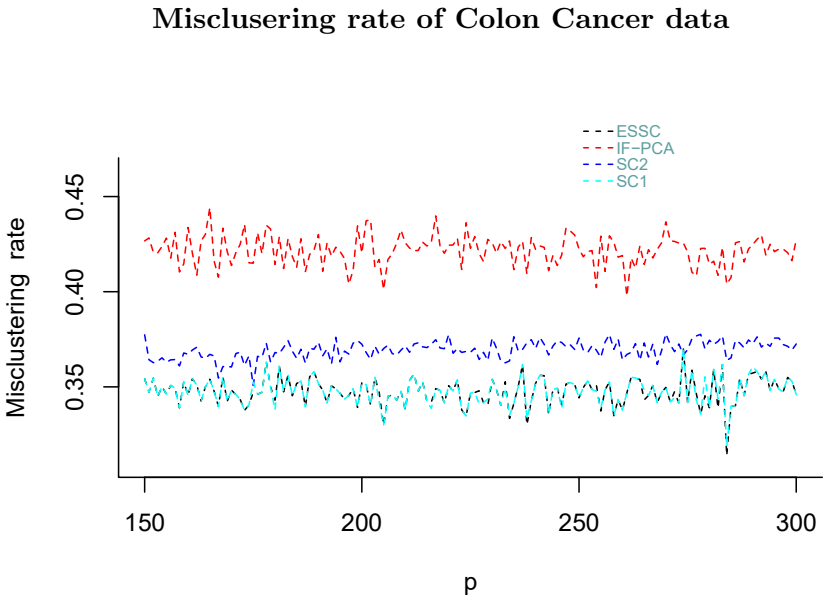

Figure 2: Misclustering rate of the Colon Cancer data vs. different feature dimension p. The red curve represents IF-PCA, the cyan curve represents SC1, the blue curve represents SC2, and the black curve represents ESSC.

From Figures 2-6, we compare the methods as follows. ESSC and SC1 work better than IF-PCA for the `Colon Cancer` and `Leukemia` data. For `Lung Cancer 1` data, ESSC has a similar misclustering rate with IF-PCA in general and outperforms the other two approaches. For `Breast Cancer` data, SC2 outperforms the other approaches, SC1 works a little better than IF-PCA, and ESSC has similar performance with SC1. For `Lung Cancer 2` data, IF-PCA has the best performance

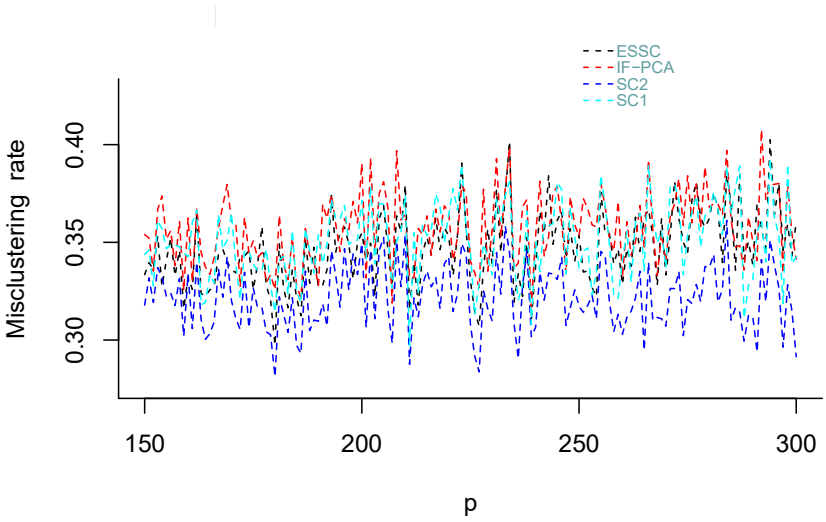

Figure 3: Misclustering rate of the Breast Cancer data vs. different feature dimension p. The red curve represents IF-PCA, the cyan curve represents SC1, the blue curve represents SC2 and the black curve represents ESSC.

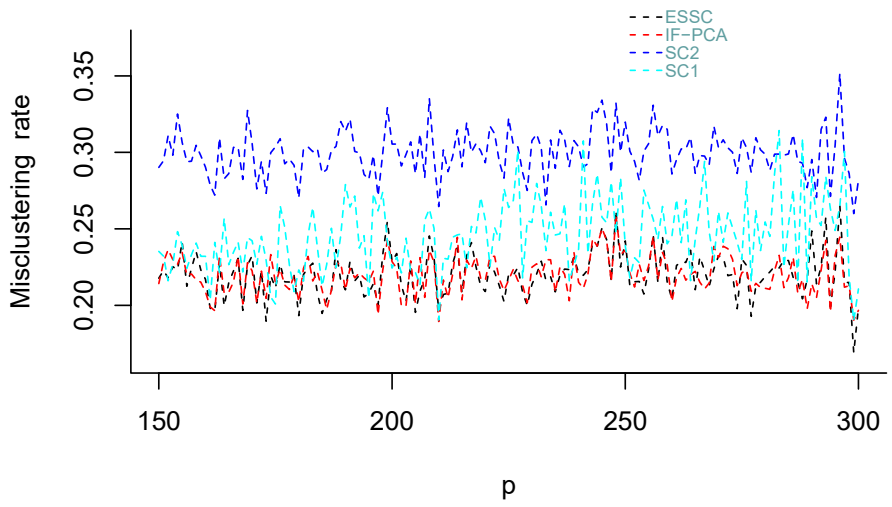

Figure 4: Misclustering rate of Lung Cancer 1 data vs. different feature dimension p. The red curve represents IF-PCA, the cyan curve represents SC1, the blue curve represents SC2 and the black curve represents ESSC.

and ESSC is the second best. Overall, ESSC belongs to the top two across all five datasets, demonstrating its efficiency and stability.

8. DISCUSSION

In this work, with a Gaussian mixture type model, we propose a theory-backed eigen selection procedures for spectral clustering. For future work, it would be interesting to study how an eigen selection procedure might help spectral clustering when a non-linear kernel is used to create an affinity matrix.

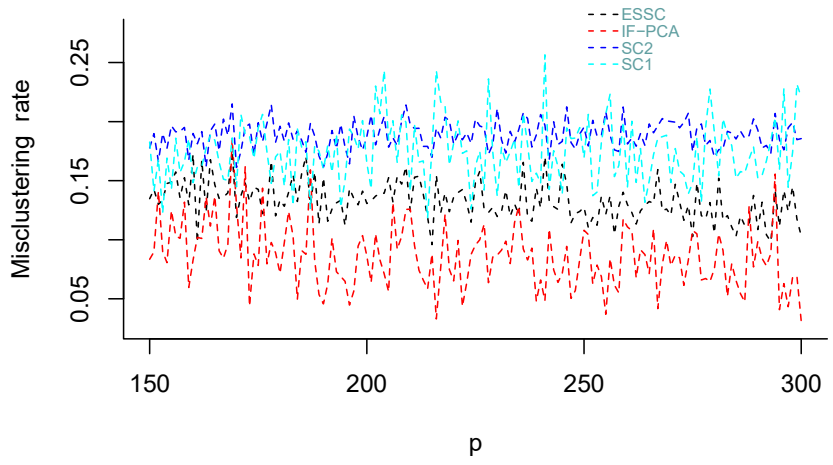

Figure 5: Misclustering rate of Lung Cancer 2 data vs. different feature dimension p. The red curve represents IF-PCA, the cyan curve represents SC1, the blue curve represents SC2 and the black curve represents ESSC.

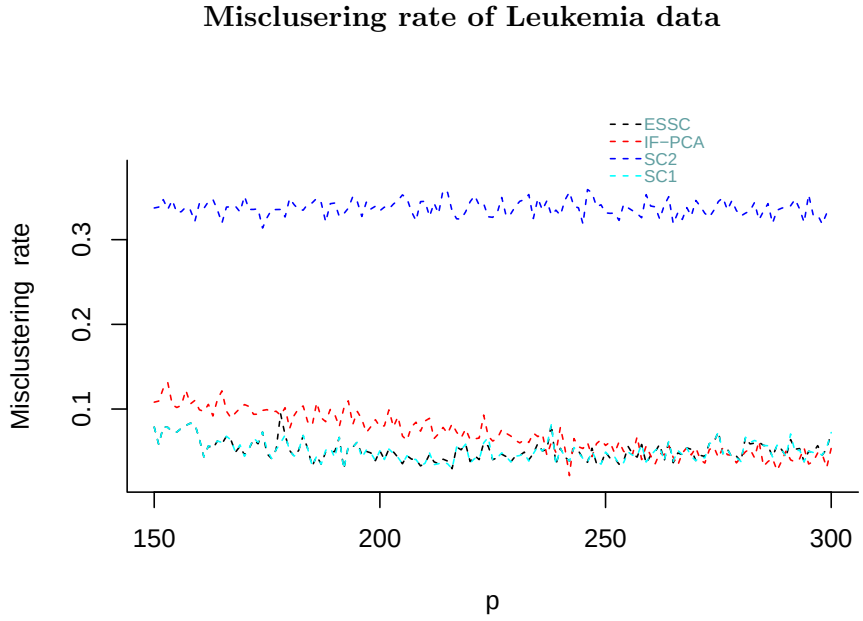

Figure 6: Misclustering rate of the Leukemia data vs. different feature dimension p. The red curve represents IF-PCA, the cyan curve represents SC1, the blue curve represents SC2 and the black curve represents ESSC.

REFERENCES

Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd edition)*. Springer-Verlag Inc, 2009. ISBN 0-387-95284-5.

G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R*. Springer Texts in Statistics. Springer New York, 2014. ISBN 9781461471370. URL https://books.google.com.hk/books?id=at1bmAEACAAJ.

Paul S Bradley, Usama M Fayyad, and Olvi L Mangasarian. Mathematical programming for data mining: Formulations and challenges. *INFORMS Journal on Computing*, 11(3):217–238, 1999.

Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. *Journal of the American Statistical Association*, 105(490):713–726, 2010.

Joe H Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244, 1963.

Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

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

Hui Zou. Classification with high-dimensional features. *WIREs: Computational Statistics*, 11: e1453, 2019.

T. Tony Cai, Zongming Ma, and Yihong Wu. Sparse pca: Optimal rates and adaptive estimation. *The Annals of Statistics*, 41(6):3074–3110, 2013. doi: 10.1214/13-AOS1178. URL https://doi.org/10.1214/13-AOS1178.

Jiashun Jin and Wanjie Wang. Influential features pca for high dimensional clustering. *The Annals of Statistics*, 44(6):2323–2359, 2016. doi: 10.1214/15-AOS1423. URL https://doi.org/10.1214/15-AOS1423.

Tao Xiang and Shaogang Gong. Spectral clustering with eigenvector selection. *Pattern Recognition*, 41(3):1012–1029, 2008.

T Tony Cai, Jing Ma, and Linjun Zhang. Chime: Clustering of high-dimensional gaussian mixtures with em algorithm and its optimality. *The Annals of Statistics*, 47(3):1234–1267, 2019.

Yao-ban Chan and Peter Hall. Using evidence of mixed populations to select variables for clustering very high-dimensional data. *Journal of the American Statistical Association*, 105(490):798–809, 2010. doi: 10.1198/jasa.2010.tm09404. URL https://doi.org/10.1198/jasa.2010.tm09404.

Martin Azizyan, Aarti Singh, and Larry Wasserman. Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In *Advances in Neural Information Processing Systems*, pages 2139–2147, 2013.

Jiashun Jin. Fast community detection by score. *The Annals of Statistics*, 43(1):57–89, 2015. doi: 10.1214/14-AOS1265.

Emmanuel Abbe, Jianqing Fan, Kaizheng Wang, and Yiqiao Zhong. Entrywise eigenvector analysis of random matrices with low expected rank. *The Annals of Statistics*, page In print, 2020+.

Jianqing Fan, Yingying Fan, Xiao Han, and Jinchi Lv. Asymptotic theory of eigenvectors for random matrices with diverging spikes. *Journal of the American Statistical Association*, 2020+. doi: 10.1080/01621459.2020.1840990.

Zhigang Bao, Xiucai Ding, and Ke Wang. Singular vector and singular subspace distribution for the matrix denoising model. *The Annals of Statistics*, page In print, 2020+.

Afonso S Bandeira, Nicolas Boumal, and Amit Singer. Tightness of the maximum likelihood semidefinite relaxation for angular synchronization. *Mathematical Programming*, 163(1-2):145–167, 2017.

Jianqing Fan, Jianhua Guo, and Shurong Zheng. Estimating number of factors by adjusted eigenvalues thresholding. *Journal of the American Statistical Association*, 0(0):1–10, 2020.

MLG ULB and Worldline. Defeatfraud: Assessment and validation of deep feature engineering and learning solutions for fraud detection. *https://www.kaggle.com/mlg-ulb/creditcardfraud*, 2018.

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

Gavin J Gordon, Roderick V Jensen, Li-Li Hsiao, Steven R Gullans, Joshua E Blumenstock, Sridhar Ramaswamy, William G Richards, David J Sugarbaker, and Raphael Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, 62(17):4963–4967, 2002.

Mohammadmahdi R Yousefi, Jianping Hua, Chao Sima, and Edward R Dougherty. Reporting bias when using real data sets to analyze classification performance. *Bioinformatics*, 26(1):68–76, 2009.