Submitted to the Annals of Statistics

DATA SPECTROSCOPY: EIGENSPACE OF CONVOLUTION OPERATORS AND CLUSTERING

By Tao Shi[*], Mikhail Belkin[†] and Bin Yu[‡]

The Ohio State University[*†] *and University of California, Berkeley*[‡]

This paper focuses on obtaining clustering information in a distribution when *iid* data are given. First, we develop theoretical results for understanding and using clustering information contained in the eigenvectors of data adjacency matrices based on a radial kernel function (with a sufficiently fast tail decay). We provide population analyses to give insights into which eigenvectors should be used and when the clustering information for the distribution can be recovered from the data. In particular, we learned that top eigenvectors do not contain all the clustering information. Second, we use heuristics from these analyses to design the *Data Spectroscopic* clustering (DaSpec) algorithm that uses properly selected top eigenvectors, determines the number of clusters, gives data labels, and provides a classification rule for future data, all based on only one eigen decomposition. Our findings not only extend and go beyond the intuitions underlying existing spectral techniques (e.g. spectral clustering and Kernel Principal Components Analysis), but also provide insights about their usability and modes of failure. Simulation studies and experiments on real world data are conducted to show the promise of our proposed data spectroscopy clustering algorithm relative to k-means and one spectral method. In particular, DaSpec seems to be able to handle unbalanced groups and recover clusters of different shapes better than competing methods.

1. Introduction. Data clustering based on eigenvectors of a proximity/affinity matrix (or its normalized version) has become popular in machine learning, computer vision and other areas. Given data $x_1, \cdots, x_n \in \mathbb{R}^d$, this family of algorithms construct a $n \times n$ affinity matrix $(K_n)_{ij} = K(x_i, x_j)/n$ based on a kernel function, such as a Gaussian kernel $K(x,y) = e^{-\frac{\|x-y\|^2}{2\omega^2}}$. Scott and Longuet-Higgins [14] proposed an algorithm that embeds data to the space spanned by the top eigenvectors of K_n, normalizes

[*] Partially supported by NASA grant NNG06GD31G

[†] Partially supported by NSF Early Career Award 0643916

[‡] Partially supported by NSF grant DMS-0605165, ARO grant W911NF-05-1-0104, NSFC grant 60628102, a grant from MSRA, and a Guggenheim Fellowship in 2006

AMS 2000 subject classifications: Primary 62H30; Secondary 68T10

Keywords and phrases: Gaussian kernel, spectral clustering, Kernel Principal Component Analysis, Support Vector Machines, unsupervised learning

the data in that space, and groups data by investigating the block structure of inner product matrix of normalized data. Perona and Freeman [11] suggested to cluster the data into two groups by directly thresholding the top eigenvector of K_n.

Eigenvectors of normalized versions of the affinity matrix had also been used to build clustering algorithms. Shi and Malik [15] connected the affinity matrix to a graph with $(K_n)_{ij}$ as the weight of its edges. They proposed a normalized cut algorithm that separates data into two groups by thresholding the second smallest generalized eigenvector of the graph Laplacian $D_n - W_n$, where $W_n = K_n - diag(K_n)$ and D_n is a diagonal matrix with $(D_n)_{ii} = \sum_j (W_n)_{ij}$. Assuming k groups, Ng, et al. [9] suggested to embed the data in the bottom k normalized eigenvectors of the normalized graph Laplacian $L_n = I_n - D_n^{-1/2} W_n D_n^{-1/2}$ then apply k-means algorithm to group the data. For more discussions on spectral clustering, we refer the reader to Weiss [19], Dhillon, et al. [3] and Luxburg [18], which provided good surveys on the scope and the history of these spectral clustering methods.

Similar to spectral clustering methods, Kernel Principal Component Analysis (Schölkopf, et al. [13]) and spectral dimensionality reduction (Belkin and Niyogi [1]) seek lower dimensional representations of the data by embedding them into the space spanned by the top eigenvectors of K_n or the bottom ones of L_n with the expectation that this embedding keeps non-linear structure of the data. Empirical observations have also been made that KPCA can sometimes capture clusters in the data. The concept of using eigenvectors of the kernel matrix is also closely connected to other kernel methods in the machine learning literature, notably Support Vector Machines, (Vapnik [17] and Schölkopf and Smola [12]), which can be viewed as fitting a linear classifier in the eigenspace of K_n.

A simple example is given here to illustrate the connections between clusters of data and top eigenvectors of the affinity matrix. The histogram of 1000 random samples from a Gaussian mixture $0.5N(2, 1^2) + 0.5N(-2, 1^2)$ is shown in the top left panel of Figure 1, where the two top eigenvectors of K_n (Gaussian kernel with $\omega = 0.3$) are plotted in the middle and lower left panels. It is clear that each eigenvector corresponds to one mixing component and it makes sense to threshold the top eigenvector or to run clustering algorithms based on the top two. Similar to the kernel matrix, its normalized versions (graph Laplacians) also connect to the clustering information in various ways. To explore these connections, different approaches such as spectral graph theory (Hagen and Kahny [5], Shi and Malik [15], Chung [2]), random walks on graphs (Melia and Shi [8]), or perturbation theory (Ng, et al. [9]) had been taken to draw similarities between the affinity matrix (or the graph Laplacian) and a block diagonal matrix that reflects the group labels. Luxburg [18] provided a good review on these approaches.

Although empirical results and theoretical studies both suggest that those top eigenvectors are related to clustering information, the effectiveness of these algorithms heavily hinge on the choices of the kernel (and its parameters), the number of the top eigenvectors used, and the number of groups assumed. As far as we know, there are no explicit theoretical results or practical guidelines on how to make these choices. More importantly, instead of tackling these questions regarding to particular data sets, it may be more fruitful to investigate the problem from a population point of view. Williams and Seeger [20] illustrated the dependence of the spectrum of K_n on the input density distribution and analyzed this connection in the content of lower rank approximation to the kernel matrix. Their work inspired our research presented in this paper.

The goal of this paper is two-fold. First we view these spectral methods from a statistical perspective, attempting to gain some insight into when and why these algorithms are expected to work well. These analyses reveal that the top eigenvectors do not always contain all the clustering information. Moreover, when the clusters are not balanced and/or the clusters have different shapes, the top eigenvectors are inadequate and redundant at the same time. That is, some top eigenvectors can correspond to the same cluster and a fixed number of top eigenvectors can miss some clusters. Hence our second goal is to devise a clustering algorithm that intelligently pick some top eigenvectors to recover more fully the clustering information even when the clusters are not balanced and possibly have different shapes.

In this paper, we concentrate on exploring the connection between $p(x)$ and the eigenvalues and eigenfunctions of the distribution-dependent convolution operator:

$$\mathcal{K}_p f(x) = \int_{\mathbb{R}^d} K(x,y) f(y) p(y) dy. \tag{1.1}$$

The kernels we consider will be positive (semi-)definite radial kernels. Such kernels can be written as $K(x,y) = k(\|x-y\|)$, where $k : [0,\infty) \to [0,\infty)$ is a decreasing function. We will use kernels with sufficiently fast tail decay, such as the Gaussian kernel or the exponential kernel $K(x,y) = e^{-\frac{\|x-y\|}{\omega}}$.

We will show that the top eigenfunctions of $\mathcal{K}_p$ may contain clustering information of the probability distribution p. To illustrate this connection (as well as the connection with the empirical version of $\mathcal{K}_p$), let us consider the example shown in Figure 1. We plot the histograms of each component and the top eigenvector of Gaussian kernel matrix defined on samples from each component in the right panels. Notice the striking similarity between the eigenvectors of kernel matrices built on the mixture and on each component. In our previous paper, Shi, et al. [16], the connection between the spectrum of the $\mathcal{K}_p$ and parameters of a Gaussian mixture distribution was used to build a "data spectroscopy" algorithm that estimates the distribution parameters through the top eigenvalues and eigenvectors of K_n. In this paper, we extend this "data spectroscopy" framework to clustering.

The paper is organized as follows. We start with basic definitions, notations, and mathematical facts of the distribution-dependent convolution operator and its spectrum in Section 2. We point out the strong connection between $\mathcal{K}_p$ and the kernel matrix K_n, which allows us to have access to the approximate spectrum of $\mathcal{K}_p$ through K_n.

In Section 3, we characterize the dependence of eigenfunctions of $\mathcal{K}_p$ on both the distribution $p(x)$ and the kernel function $K(\cdot,\cdot)$. We show that the eigenfunctions of $\mathcal{K}_p$ decay to zero at the tails of the distribution $p(x)$ and how fast they decay depend on both the tail decay rate of $p(x)$ and that of the kernel $K(\cdot,\cdot)$. For distributions with only one high density component, we provide some theoretical analysis and discuss two examples where the exact form of the eigenfunctions of $\mathcal{K}_p$ can be obtained. We also discuss the case when the distribution is concentrated on or around a curve in $\mathbb{R}^d$.

In Section 4, we consider the case when the distribution p contains several separate high-density components. Through classical results of the perturbation theory, we show that the top eigenfunctions of $\mathcal{K}_p$ are approximated by the top eigenfunctions of the corresponding operators defined on some of those components. However, not every component will contribute to the top few eigenfunctions of $\mathcal{K}_p$ as the corresponding eigenvalues are determined by the size and configuration of the component. Based on this key property, we show why the top eigenvectors of the kernel matrix may or may not preserve all clustering information, which explains some empirical observations of certain spectral clustering methods.

In Section 5, we utilize our theoretical results to construct a *Data Spectroscopic* clustering (DaSpec) algorithm that estimates the number of groups data-dependently, assigns labels to each observation, and provides a classification rule for unobserved data, all based on the same eigen decomposition. Data-dependent choices of algorithm parameters are also discussed. In Section 6, the proposed DaSpec algorithm is tested on two simulations and the USPS post code data against commonly used k-means and spectral clustering algorithms. In all three situations, the DaSpec algorithm provides favorable results even when other two algorithms are provided with reasonable group numbers. We conclude the paper in Section 7.

2. Notations and Mathematical Preliminaries.

2.1. *Distribution-dependent Convolution Operator.* Given a probability distribution $p(x)$ on $\mathbb{R}^d$, we define $L_p^2(\mathbb{R}^d)$ to be the space of square integrable functions, $f \in L_p^2(\mathbb{R}^d)$ if $\int_{\mathbb{R}^d} f^2(x)p(x)dx < \infty$, and the space is equipped with an inner product $\langle f, g\rangle = \int_{\mathbb{R}^d} f(x)g(x)p(x)dx$. Given a kernel (symmetric function of two variables) $K(x,y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, Eq. (1.1) defines the corresponding integral operator $\mathcal{K}_p$. Recall that an eigenfunction $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ and the corresponding eigenvalue λ of $\mathcal{K}_p$ are defined by the following equations:

$$\mathcal{K}_p \phi = \lambda \phi. \tag{2.1}$$

If the kernel satisfies the condition

$$\int\int K^2(x,y)p(x)p(y)dxdy < \infty, \tag{2.2}$$

the corresponding operator $\mathcal{K}_p$ is a trace class operator, which, in turn, implies that it is compact and carries a discrete spectrum.

In this paper we will only consider the case when a positive semi-definite kernel $K(x,y)$ and $p(x)$ generate a trace class operator $\mathcal{K}_p$, so that it has only countable non-negative eigenvalues $\lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \ldots \geq 0$. Moreover, there is a corresponding orthonormal basis in L_p^2 of eigenfunctions ϕ_i that satisfies Eq. (2.1). The dependence of the eigenvalues and eigenfunctions of $\mathcal{K}_p$ on p will be one of the main foci of our paper. We want to emphasize that the eigenfunction ϕ is uniquely defined not only on the support of $p(x)$, but on every point $x \in \mathbb{R}^d$ through $\phi(x) = \frac{1}{\lambda}\int K(x,y)p(y)\phi(y)dy$, assuming that the kernel function K is defined on $\mathbb{R}^d \times \mathbb{R}^d$.

2.2. *Kernel Matrix.* Let $x_1, \ldots, x_n$ be an *i.i.d.* sample drawn from a distribution $p(x)$. The corresponding empirical operator $\mathcal{K}_{p_n}$ is defined as

$$\mathcal{K}_{p_n} f(x) = \frac{1}{n}\sum_{i=1}^{n} K(x_i, x) f(x_i).$$

This operator is closely related to the $n \times n$ kernel matrix K_n, where

$$(K_n)_{ij} = K(x_i, x_j)/n.$$

Specifically, the eigenvalues of $\mathcal{K}_{p_n}$ are the same as those of K_n and an eigenfunction ϕ, with an eigenvalue $\lambda \neq 0$ of $\mathcal{K}_{p_n}$, is connected with the corresponding eigenvector $\boldsymbol{v} = (\boldsymbol{v}(x_1), \ldots, \boldsymbol{v}(x_n))'$ of K_n by

$$\phi(x) = \frac{1}{n\lambda}\sum_{i=1}^{n} \boldsymbol{v}(x_i) K(x_i, x) \qquad \forall x \in \mathbb{R}^d.$$

It is easy to verify that $\mathcal{K}_{p_n}\phi = \lambda\phi$. Thus values of ϕ at locations $x_1, \ldots, x_n$ coincide with the corresponding entries of the eigenvector $\boldsymbol{v}$. However, unlike $\boldsymbol{v}$, ϕ is defined everywhere in $\mathbb{R}^d$. For the spectrum of $\mathcal{K}_{p_n}$ and K_n, the only difference is that the spectrum of $\mathcal{K}_{p_n}$ contains 0 with infinite multiplicity. The corresponding eigenspace includes all functions vanishing on the sample points.

It is well-known (e.g. Koltchinskii and Giné [6]) that, under mild conditions, the eigenvectors and eigenvalues of K_n converge to eigenfunctions and eigenvalues of $\mathcal{K}_p$ as $n \to \infty$. Therefore, we expect the properties of the top eigenfunctions and eigenvalues of $\mathcal{K}_p$ to also hold for K_n, assuming that n is reasonably large.

3. Spectral Properties of a Single Component.

In this section we will give some properties and examples for the case when our distribution p consists of one (high-density) component. Through a theorem and its corollary, we obtain an important property of the eigenfunctions of $\mathcal{K}_p$ showing a fast decay away from the majority of masses of the component if the tails of K and p have a fast decay. Another theorem offers the important property of the leading eigenfunction that it has no sign change and multiplicity one. The section ends with detailed examples to illustrate the important properties from those theorems and corollary.

Theorem 1 (Tail decay property of eigenfunctions). *An eigenfunction ϕ with the corresponding eigenvalue $\lambda > 0$ of $\mathcal{K}_p$ satisfies*

$$|\phi(x)| \leq \frac{1}{\lambda}\sqrt{\int K^2(x,y)p(y)dy}.$$

Proof: By definition,

$$\lambda\phi(x) = \int K(x,y)\phi(y)p(y)dy.$$

For simplicity, we assume that the density exists, but the same argument can be made for a general probability distribution. By the Cauchy-Schwartz inequality, we see that

$$\begin{aligned}
\lambda|\phi(x)| &\leq \int K(x,y)|\phi(y)|p(y)dy = \int \left[K(x,y)p^{1/2}(y)\right]\left[|\phi(y)|p^{1/2}(y)\right]dy \\
&\leq \sqrt{\int K^2(x,y)p(y)dy}\sqrt{\int \phi^2(y)p(y)dy} = \sqrt{\int K^2(x,y)p(y)dy}.
\end{aligned}$$

The conclusion follows. □

We see that the "tails" of eigenfunctions of $\mathcal{K}_p$ decay to zero and that the decay rate depends on the tail behaviors of both the kernel K and the distribution p. This observation will be useful to separate high-density areas in the case of p having several components. Actually, we have the following corollary immediately:

COROLLARY 1. *Let $K(x,y) = k(\|x-y\|)$. Assume that p is supported on a compact set $D \subset \mathbb{R}^d$. Then*

$$|\phi(x)| \leq \frac{k\left(\operatorname{dist}(x,D)\right)}{\lambda}$$

where $\operatorname{dist}(x,D) = \inf_{y \in D} \|x-y\|$.

Proof: Follows from Theorem 1 and the fact that $k(\cdot)$ is a decreasing function. □

Next we give an important property of the top (corresponding to the largest eigenvalue) eigenfunction.

THEOREM 2 (Top eigenfunction). *Let $K(x,y)$ be a positive semi-definite kernel with full support on $\mathbb{R}^d$,. The top eigenfunction $\phi_0(x)$ of the convolution operator $\mathcal{K}_p$*

1. *is the only eigenfunction with no sign change on $\mathbb{R}^d$;*
2. *has multiplicity one;*
3. *is non-zero on the support of p.*

The proof is given in the appendix and these properties will be used later when we propose our data spectroscopic clustering algorithm in Section 6. To illustrate these theoretical results we now study some concrete examples.

Example 1: Gaussian kernel, Gaussian density

Let us start with the univariate Gaussian case where both the probability distribution p is $N(\mu, \sigma^2)$ and the kernel function is also Gaussian. The following proposition (Shi, et al. [16]) about the Gaussian Convolution Operator $\mathcal{K}_p$, is a slightly refined version of a result in Zhu, et al. [21].

PROPOSITION 1. *For $p \sim N(\mu, \sigma^2)$ and a Gaussian kernel $K(x,y) = e^{-\frac{(x-y)^2}{2\omega^2}}$, let $\beta = 2\sigma^2/\omega^2$ and let $H_i(x)$ be the i-th order Hermite polynomial.*

Then eigenvalues and eigenfunctions of $\mathcal{K}_p$ for $i = 0, 1, \cdots$ are given by

$$\lambda_i = \sqrt{\frac{2}{(1+\beta+\sqrt{1+2\beta})}} \left(\frac{\beta}{1+\beta+\sqrt{1+2\beta}}\right)^i, \tag{3.1}$$

$$\phi_i(x) = \frac{(1+2\beta)^{1/8}}{\sqrt{2^i i!}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\frac{\sqrt{1+2\beta}-1}{2}\right) H_i\left(\left(\frac{1}{4}+\frac{\beta}{2}\right)^{\frac{1}{4}} \frac{x-\mu}{\sigma}\right). \tag{3.2}$$

Here H_k is the k-th order Hermite Polynomial:

$$\begin{aligned} H_0(x) &= 1, & H_1(x) &= 2x, \\ H_2(x) &= 4x^2 - 2, & H_3(x) &= 8x^3 - 12x. \end{aligned}$$

Clearly from the explicit expression and expected from Theorem 2, ϕ_0 is the only positive eigenfunction of $\mathcal{K}_p$. We note that each eigenfunction ϕ_i decays quickly (as it is a Gaussian multiplied by a polynomial) away from the mean μ of the probability distribution p. We also see that the eigenvalues of $\mathcal{K}_p$ decay exponentially with the rate dependent on the bandwidth of the Gaussian kernel ω and the variance of the probability distribution σ^2. These observations can be easily generalized to the multivariate case, see Shi, et al. [16].

Example 2: Exponential kernel, uniform distribution on an interval.

To give another concrete example, consider the exponential kernel $K(x, y) = \exp(-\frac{|x-y|}{\omega})$ for the uniform distribution on the interval $[-1, 1] \subset \mathbb{R}$. In Diaconis, et al. [4] it was shown that the eigenfunctions of this kernel can be written as $\cos(bx)$ or $\sin(bx)$ inside the interval $[-1, 1]$ for appropriately chosen values of b and decay exponentially away from it. The top eigenfunction can be written explicitly as follows:

$$\phi(x) = \frac{1}{\lambda}\int_{[-1,1]} e^{-\frac{|x-y|}{\omega}} \cos(by)\, dy, \qquad \forall x \in \mathbb{R},$$

where λ is the corresponding eigenvalue. Figure 2 illustrates an example of this behavior, for $\omega = 0.5$.

Example 3: A curve in $\mathbb{R}^d$. We now give a brief informal discussion of the important case when our probability distribution is concentrated on or around a low-dimensional submanifold of an (potentially high-dimensional) ambient space. The simplest example of this setting is a Gaussian distribution, which can be viewed as a zero-dimensional manifold (the mean of the distribution) plus noise.

A more interesting example of a manifold is a curve in $\mathbb{R}^d$. We observe that such data is generated by any time-dependent smooth deterministic process, whose parameters depend continuously on time t. Let $\psi(t) : [0,1] \to \mathbb{R}^d$ be such a curve. Consider a restriction of the kernel $\mathcal{K}_p$ to ψ. Let $x, y \in \psi$ and let $d(x,y)$ be the geodesic distance along the curve. It can be shown that $d(x,y) = \|x-y\| + O(\|x-y\|^3)$, when x, y are close, with the remainder term depending on how the curve is embedded in $\mathbb{R}^d$. Therefore, we see that if the kernel $\mathcal{K}_p$ is a sufficiently local radial basis kernel, the restriction of $\mathcal{K}_p$ to ψ is a perturbation of $\mathcal{K}_p$ in a one-dimensional case. For the exponential kernel, the one-dimensional kernel can be written explicitly (see Example 2) and we have an approximation to the kernel on the manifold with a decay off the manifold (assuming that the kernel is a decreasing function of the distance). For the Gaussian kernel similar extension holds, although no explicit formula can be easily obtained.

The behaviors of the top eigenfunction of the Gaussian and exponential kernel respectively are demonstrated in Figure 3. The exponential kernel is the bottom left panel. We see that the behavior of the eigenfunction is generally consistent with the top eigenfunction of the exponential kernel on $[-1,1]$ shown in Figure 2. We see that the Gaussian kernel (top left panel) has similar behaviors but produces level lines more consistent with the data distribution, which may be preferable in practice. Finally we observe that the addition of small noise (right top and bottom panels) does not significantly change the eigenfunctions.

4. Spectral Properties of Mixture Distributions.

In this section, we study the spectrum of $\mathcal{K}_p$ defined on a mixture distribution

$$p(x) = \sum_{g=1}^{G} \pi^g p^g(x),$$

which is a commonly used model in clustering and classification. The main result is that, if the kernel has a sufficiently fast tail decay, each of the top eigenfunctions of $\mathcal{K}_p$ connects directly to one of the separable mixing components. However, some top eigenfunctions can correspond to the same component and a fixed number of top eigenfunctions may miss some components.

We start by revisiting the mixture Gaussian example given in Figure 1. For Gaussian kernel matrices K_n, K_n^1, and K_n^2 ($\omega = 0.3$) constructed on samples from $0.5N(2,1^2) + 0.5N(-2,1^2)$, $N(2,1^2)$ and $N(-2,1^2)$ respectively, the top eigenvectors of K_n are nearly identical to the top eigenvectors of K_n^1 or K_n^2. From the point of view of the operator theory, it is easy to understand this phenomenon: the top eigenfunctions of an operator defined on each mixing component are approximate eigenfunctions of the operator defined on the mixture distribution. To be explicit, let us consider the Gaussian convolution operator $\mathcal{K}_p$ defined by $p(x) = \pi^1 p^1 + \pi^2 p^2$, with Gaussian components $p^1 = N(\mu^1, (\sigma^1)^2)$ and $p^2 = N(\mu^2, (\sigma^2)^2)$ and the Gaussian kernel $K(x, y)$ with bandwidth ω. The corresponding convolution operators are $\mathcal{K}_{p^1}$ and $\mathcal{K}_{p^2}$ and $\mathcal{K}_p = \pi^1\mathcal{K}_{p^1} + \pi^2\mathcal{K}_{p^2}$ respectively.

Consider an eigenfunction $\phi^1(x)$ of $\mathcal{K}_{p^1}$ with the corresponding eigenvalue λ^1, $\mathcal{K}_{p^1}\phi^1(x) = \lambda^1\phi^1(x)$. We have

$$\mathcal{K}_p\phi^1(x) = \pi^1\lambda^1\phi^1(x) + \pi^2 \int_{-\infty}^{\infty} K(x, y)\phi^1(y)p^2(y)dy.$$

As we have shown in Proposition 1 in Section 3, any eigenfunction $\phi^1(x)$ of $\mathcal{K}_{p^1}$ is centered at μ^1 and decays exponentially away from μ^1. Therefore, assuming the separation $|\mu^1 - \mu^2|$ is large enough, the second summand $\pi^2 \int K(x, y)\phi^1(x)p^2(x)dx$ is close to 0 everywhere and hence $\phi^1(x)$ is an approximate eigenfunction of $\mathcal{K}_p$. Shi et al. [16] utilized these findings to construct a data spectroscopic algorithm to estimate mixture Gaussian distributions. In the next section, we will show that this approximation result holds for general mixture distributions beyond Gaussian components, and this leads to a Data Spectroscopic clustering algorithm described in Section 5.

4.1. *Perturbation Analysis.*

For a positive semi-definite kernel $K(\cdot, \cdot)$ and $p(x) = \sum_{g=1}^{G} \pi^g p^g(x)$ on $\mathbb{R}^d$, we now study the connection between the top eigenvalues and eigenfunctions of $\mathcal{K}_p$ to those of each $\mathcal{K}_{p^g}$. We note that the results shown here only require the operator $\mathcal{K}_p$ possessing a discrete spectrum. Without loss of generality, let us start with a mixture of two components, e.g. $p = \pi^1 p^1 + \pi^2 p^2$ and $\pi^1 + \pi^2 = 1$. We have the following theorem regarding the top eigenvalue λ_0 of $\mathcal{K}_p$.

THEOREM 3 (Top eigenvalue of mixture distribution). *Let p^1 and p^2 be probability distributions on $\mathbb{R}^d$ and define their mixture as $p = \pi^1 p^1 + \pi^2 p^2$ with $\pi^1 + \pi^2 = 1$. Given a positive semi-definite kernel K, denote the top eigenvalue of $\mathcal{K}_p$, $\mathcal{K}_{p^1}$ and $\mathcal{K}_{p^2}$ as λ_0, λ_0^1 and λ_0^2 respectively. Then λ_0 satisfies*

$$max(\pi^1\lambda_0^1, \pi^2\lambda_0^2) \le \lambda_0 \le max(\pi^1\lambda_0^1, \pi^2\lambda_0^2) + r,$$

where

$$r = \left(\pi^1\pi^2 \iint [K(x, y)]^2 p^1(x)p^2(y)dxdy\right)^{1/2}. \tag{4.1}$$

The proof is given in the appendix. As illustrated in Figure 4, the value of r in Eq [4.1] is small when p^1 and p^2 do not overlap much. Meanwhile, the size of r is also affected by how fast $K(x, y)$ approaches zero as $\|x - y\|$ increases. When the separation condition is satisfied, the top eigenvalue of $\mathcal{K}_p$ is close to the larger one of $\pi^1\lambda_0^1$ and $\pi^2\lambda_0^2$. Without loss of generality, we assume $\pi^1\lambda_0^1 > \pi^2\lambda_0^2$ in the rest of this section.

The next lemma is a general perturbation result that deals with eigenfunctions of $\mathcal{K}_p$. The empirical (matrix) version of this lemma appeared in Diaconis et al. [4] and more general results can be traced back to Parlett [10].

Lemma 1. *Consider an operator $\mathcal{K}_p$ with a discrete spectrum $\lambda_0 \geq \lambda_1 \geq \lambda_2 \geq \cdots$. If*

$$\|\mathcal{K}_p f - \lambda f\|_{L_p^2} \leq \epsilon$$

for some λ, $\epsilon > 0$, and $f \in L_p^2$, then $\mathcal{K}_p$ has an eigenvalue λ_k such that $|\lambda_k - \lambda| \leq \epsilon$.
If we further assume that

$$s = \min_{i:\lambda_i \neq \lambda_k} |\lambda_i - \lambda_k| > \epsilon$$

then $\mathcal{K}_p$ has an eigenfunction f_k such that $\mathcal{K}_p f_k = \lambda_k f_k$ and $\|f - f_k\|_{L_p^2} \leq \frac{\epsilon}{s-\epsilon}$.

This lemma shows that a constant λ must be "close" to an eigenvalue λ_k of $\mathcal{K}_p$ if the operator "almost" projects a function f to λf. Moreover, the function f must be "close" to an eigenfunction of $\mathcal{K}_p$ if the distance between $\mathcal{K}_p f$ and λf is smaller than the eigen-gaps between λ_k and other eigenvalues of $\mathcal{K}_p$. The reader is referred to Diaconis et al. [4] for a detailed proof, and more refined results may be found in Parlett [10].

We are now ready to state the perturbation result for the top eigenfunction of $\mathcal{K}_p$ defined on the mixture. Given the facts that $|\lambda_0 - \pi^1\lambda_0^1| \leq r$ and

$$\mathcal{K}_p\phi_0^1 = \pi^1\mathcal{K}_{p^1}\phi_0^1 + \pi^2\mathcal{K}_{p^2}\phi_0^1 = (\pi^1\lambda_0^1)\phi_0^1 + \pi^2\,\mathcal{K}_{p^2}\phi_0^1,$$

Lemma 1 indicates that ϕ_0^1 is close to ϕ_0 if $\|\pi^2\mathcal{K}_{p^2}\phi_0^1\|_{L_p^2}$ is small enough. To be explicit, we have the following corollary.

Corollary 2 (Top eigenfunction of mixture distribution). *For a given semi-positive definite kernel $K(\cdot,\cdot)$, consider a convolution operator defined as $\mathcal{K}_p(\cdot) = \int K(x,\cdot)f(x)p(x)dx$. Let p^1 and p^2 be two probability distributions and $p = \pi^1p^1 + \pi^2p^2$. Denote the top eigenvalues of $\mathcal{K}_{p^1}$ and $\mathcal{K}_{p^2}$ as λ_0^1 and*

λ_0^2 respectively (assuming $\pi^1\lambda_0^1 > \pi^2\lambda_0^2$) and define $t = \lambda_0 - \lambda_1$, the eigen-gap of $\mathcal{K}_p$. If the constant r defined in Eq.(4.1) satisfies $r < t$, and

$$\left\| \pi^2 \int_{\mathbb{R}^d} K(x,y)\phi_0^1(y)p^2(y)dy \right\|_{L_p^2} \leq \epsilon \tag{4.2}$$

such that $\epsilon + r < t$, then λ_0^1 is close to the top eigenvalue (λ_0) of $\mathcal{K}_p$,

$$|\pi^1\lambda_0^1 - \lambda_0| \leq \epsilon,$$

and $\phi^1\phi_0^1$ is close to the top eigenfunction (ϕ_0) of $\mathcal{K}_p$ in the sense:

$$\|\phi_0^1 - \phi_0\|_{L_p^2} \leq \frac{\epsilon}{t-\epsilon}. \tag{4.3}$$

Since Theorem 3 leads to $|\lambda_0^1 - \lambda_0| \leq r$ and Lemma 1 suggests $|\lambda_0^1 - \lambda_k| \leq \epsilon$ for some k, the condition $r + \epsilon < t = \lambda_0 - \lambda_1$ guarantees that ϕ_0 as the only possible choice for ϕ_0^1 to be close to. Therefore, ϕ_0^1 is approximately the top eigenfunction of $\mathcal{K}_p$. Therefore, condition (4.2) is another separation condition required to connect the top eigenfunction of $\mathcal{K}_p$ to the top ones of $\mathcal{K}_{p^g}$.

4.2. *Top Spectrum of $\mathcal{K}_p$ on Mixture Distributions.* For the $\mathcal{K}_p$ defined on the mixture distribution ($p = \pi^1p^1 + \pi^2p^2$), we now extend the perturbation results on the top eigenfunction to other top ones. We know from Theorem 1 that $|\phi_0^1(x)|$ decay exponentially as x get away from the majority mass of $p^1(x)$. In the case that $p^2(x)$ has little overlap with $\phi_0^1(x)$, $|\phi_0^1(x)|p^2(x)$ will be close to zero everywhere, which makes condition (4.2) satisfied. This condition also holds for other top eigenfunctions of $\mathcal{K}_{p^1}$ since they also decay to zero quickly as x moves away from the majority mass of p^1. The same argument applies to the top eigenfunctions of $\mathcal{K}_{p^2}$ as well.

With a high quality agreement between (λ_0, ϕ_0) and $(\pi^1\lambda_0^1, \phi_0^1)$, we can also derive the conditions under which the second eigenvalue of $\mathcal{K}_p$ is approximately $max(\pi^1\lambda_1^1, \pi^2\lambda_0^2)$ by working with a new kernel $K^{new} = K(x,y) - \lambda_0\phi_0(x)\phi_0(y)$. Then we can show that ϕ_1 of $\mathcal{K}_p$ is close to ϕ_1^1 or ϕ_0^2, depending on which one corresponds to $\max(\pi^1\lambda_1^1, \pi^2\lambda_0^2)$. By sequentially applying the same argument, we arrive at the following important property:

Mixture property of top spectrum: For a convolution operator $\mathcal{K}_p$ with a fast tail decay kernel and enough separations between components of a mixture distribution $p(x) = \sum_{g=1}^{G} \pi^g p^g(x)$, the top eigenfunctions ϕ_j of $\mathcal{K}_p$ are approximately chosen from the top ones (ϕ_i^g) of $\mathcal{K}_{p^g}$, $i = 0, 1, \cdots$, and

$g = 1, \cdots, G$. The ordering of the eigenfunctions are determined by *mixture magnitudes* $\pi^g \lambda_i^g$.

Along with the mixture property, we note the following useful facts about the top spectrum of $\mathcal{K}_p$ defined on mixture distributions:

1. We gain access to approximate the top eigenfunctions of $\mathcal{K}_{p^g}$ through those of $\mathcal{K}_p$ when enough separations among components of p exist.
2. The sizes of the mixture magnitudes $\pi^g \lambda_i^g$ determine the ordering of $(\pi^g \lambda_i^g, \phi_i^g)$ in the top spectrum of $\mathcal{K}_p$. In other words, the top eigenfunction of $\mathcal{K}_{p^g}$ with a small mixing weight π^g or small λ's may come into the spectrum of $\mathcal{K}_p$ far below the top eigenvalues.
3. We point out that the separable conditions in Theorem 3 and Corollary 2 are mainly based on the overlap of the mixture components, but not on their shapes or parametric forms. Therefore, clustering methods based on spectral information are able to deal with more general problems beyond the traditional mixture models based on a parametric family, such as mixture Gaussians or mixture of exponential families.
4. The separation between components p^{g_i} and p^{g_j} does not need to be as perfect as none-overlapping. A good separation is achieved as long as $\left(\pi^i \pi^j \iint [K(x,y)]^2 p^i(x) p^j(y) dx dy\right)^{1/2}$ is small relative to the eigengap (Corollary 2). As demonstrated in the example given in Figure 1, the approximations of eigenfunctions hold well even if the components have significant overlaps.

When data are collected *i.i.d.* from the mixture distribution, we expect the top eigenvalues and eigenfunctions of $\mathcal{K}_p$ are well approximated by those of the empirical operator $\mathcal{K}_{p_n}$. As we also discussed in Section 2.2, the eigenvalues of $\mathcal{K}_{p_n}$ is the same as those of the kernel matrix K_n and the eigenfunctions of $\mathcal{K}_{p_n}$ coincident with the eigenvectors of K_n on the sampled points. Therefore, assuming good approximation of $\mathcal{K}_{p_n}$ to $\mathcal{K}_p$, the eigenvalues and eigenvectors of K_n provide us with access to the spectrum of $\mathcal{K}_p$.

This understanding sheds light on the algorithms proposed in Scott and Longuet-Higgins [14] and Perona and Freeman [11], in which the top (several) eigenvectors of K_n are used for clustering. Given good approximation of K_n to $\mathcal{K}_p$, we see from **Fact 2** that while top eigenvectors may contain clustering information, smaller or less compact groups may not be identified using just the very top part of the spectrum, More eigenvectors need to be investigated to see those clusters. On the other hand, information in the top few eigenvectors may also be redundant for clustering, as some of these eigenvectors may represent the same group. This observation is also supported by our experimental results in Section 6.

5. A Data Spectroscopic Clustering (DaSpec) Algorithm.

In this section, we propose a Data Spectroscopic clustering (DaSpec) algorithm based on our theoretical analyses on the spectrum of $\mathcal{K}_p$ relative to that of the clustering components of a distribution. We chose the commonly used Gaussian kernel in the proposed algorithm, but it may be replaced by other positive semi-definite radial kernels with a fast tail decay rate.

5.1. *Justification and the DaSpec Algorithm.*

Because of the **Mixture property of top spectrum** of $\mathcal{K}_p$ defined on mixture distributions, we have access to approximate eigenfunctions of $\mathcal{K}_{p^g}$ through those of $\mathcal{K}_p$ when each mixing component has enough separation from others. Among the eigenfunctions of each $\mathcal{K}_{p^g}$, we know from Theorem 2 that the top one is the only one with no sign changes. Given the nature of the approximation of spectrum of $\mathcal{K}_{p^g}$ to that of $\mathcal{K}_p$, we expect that there is one and only one eigenfunction with no sign changes over a certain small threshold ϵ on $|\phi(x)|$. Therefore, the number of separable components of p is indicated by the number of eigenfunctions $\phi(x)$'s of $\mathcal{K}_p$ with no sign changes after thresholding on $|\phi(x)|$.

Meanwhile, the eigenfunctions of each component decay quickly to zeros at the tail of its distribution. At a given location x in the high density area of a particular component, which is at the tails of other components, we expect the eigenfunctions from all other components to be close to zero. Among the top eigenfunction $|\phi_0^g(x)|$ of $\mathcal{K}_{p^g}$ defined on each component p^g, $g = 1, \ldots, G$, the group identity of x is tied to the eigenfunction that has the largest absolute value in $|\phi_0^g(x)|$. Combining this observation with previous discussions on the approximation of K_n to $\mathcal{K}_p$, we propose the following clustering algorithm.

Data Spectroscopic clustering (DaSpec) Algorithm
Input: Data $x_1, \ldots, x_n \in \mathbb{R}^d$.
Parameters: Gaussian kernel bandwith $\omega > 0$, thresholds $\epsilon_j > 0$ for $j = 1, 2, \ldots, n$
Output: Estimated Number $\hat{G}$ of clustering components and a cluster label for each data point.

Step 1. Constructing the Gaussian kernel matrix K_n:

$$(K_n)_{ij} = \frac{1}{n} e^{-\frac{\|x_i - x_j\|^2}{2\omega^2}}, \quad i,j = 1,\ldots,n$$

Compute its (top) eigenvalues $\lambda_1, \lambda_2, \ldots$ and eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots$

Step 2. Estimating the number of clusters $\hat{G}$:
Identify all eigenvectors $\boldsymbol{v}_j$ that have no sign changes up to precision ϵ_j
(We say that a vector $\boldsymbol{e} = (e_1, \ldots, e_n)$ has no sign changes up to ϵ
if either $\forall_i \;\; e_i > -\epsilon$ or $\forall_i \;\; e_i < \epsilon$)
Estimate the number of groups by $\hat{G}$, the number of such eigenvectors.
Denote these eigenvectors and the corresponding eigenvalues by
$\boldsymbol{v}_0^1, \boldsymbol{v}_0^2, \ldots, \boldsymbol{v}_0^{\hat{G}}$ and $\lambda_0^1, \lambda_0^2, \ldots, \lambda_0^{\hat{G}}$ respectively.

Step 3. Assigning a cluster label to each data point:
For a data point x_i, assign its label as

$$\text{argmax}_g \{abs(\boldsymbol{v}_0^g(x_i)) : \quad g = 1, 2, \cdots, \hat{G}\}$$

One important feature of our algorithm is that little adjustment is needed to classify an unobserved data point x. Thanks to the connection between the eigenvector $\boldsymbol{v}$ of K_n and the eigenfunction ϕ of the empirical operator $\mathcal{K}_{p_n}$, we can compute the eigenfunction ϕ_0^g corresponding to $\boldsymbol{v}_0^g$ by

$$\phi_0^g(x) = \frac{1}{\lambda} \sum_{i=1}^{n} \boldsymbol{v}_0^g(x_i) K(x_i, x) \qquad \forall x \in \mathbb{R}^d.$$

Therefore, **Step 3** of the algorithm can be readily applied to any x by replacing $\boldsymbol{v}_0^g(x_i)$ with $\phi_0^g(x)$. So the algorithm output can server as a clustering rule that separates not only the data, but also the underline distribution, which is aligned with the motivation behind our Data Spectroscopy framework: learning properties of a distribution though the spectrum of $\mathcal{K}_{p_n}$.

5.2. *Data-dependent Parameter Specification.* Following the justification of our DaSpec algorithm, we provide some guidelines for choosing algorithm parameters in practice.

Gaussian kernel bandwidth ω: The Gaussian kernel bandwidth ω controls the gaps of between eigenvalues and the tail decay rate of the eigenfunctions. When ω is too large, the tails of eigenfunctions may not decay fast enough to make condition (4.2) in Corollary 2 hold. In principle, we want to have ω small enough to keep eigenfunctions decaying fast at the tail. On the other side, if ω is too small, the eigen-gaps may vanish, in which case each data point will end up as a separate group.

In practice, we suggest to select ω as the smallest value that keeps most data points (say 95% of them) having a certain number (5% of sample size) of neighbors within the "range" of the kernel. For a d-dimensional Gaussian kernel with a bandwidth ω, we define the range of the kernel as the length l that makes $P(\|X\| < l) = 95\%$, where $X \in \mathbb{R}^d$ follows $N(0, \omega^2 I)$. In this case, the range $l = \omega\sqrt{95\% \text{ quantile of } \chi^2_d}$, since $\|X\|^2/\omega^2$ follows a χ^2 distribution with d degrees of freedom.

Given data $x_1, \ldots, x_n$ or their pairwise L^2 distance $d(x_i, x_j)$, we may find ω that satisfies the above criteria by first calculating $q_i = 5\%$ quantile of $\{d(x_i, x_j),\ j = 1, \ldots, n\}$ for each $i = 1, \ldots, n$, then taking

$$\omega = \frac{95\% \text{ quantile of } \{q_1, \ldots, q_n\}}{\sqrt{95\% \text{ quantile of } \chi^2_d}}. \tag{5.1}$$

As shown in the simulation studies in Section 6, this particular choice of ω works well in low dimensional case. For high dimensional data generated from a lower dimensional structure, such as an m-manifold, our procedure usually leads to an ω that is too small, since the quantile of χ^2_d is larger than the corresponding quantile of χ^2_m that should be used in (5.1). Therefore, we suggest starting with ω defined in (5.1) and trying some neighboring ones to see if the results get improved, which may be based on some labeled data, expert opinions, data visualization or trade-off of the between and within cluster distances.

Threshold ϵ_j: When identifying the eigenvectors with no sign changes of each group in **Step 2**, a threshold ϵ_j is included to deal with the perturbation introduced by other groups. For the top eigenvector $\boldsymbol{v}(x)$ of a particular group, The values $(|\boldsymbol{v}(x)|)$ at x's of this group are much larger than those of other groups, even after the perturbation. Given $\|\boldsymbol{v}_j\|^2 = \sum_i \boldsymbol{v}_j(x_i)^2 = 1$ and the absolute values of its entries decrease quickly (exponentially) away from $\max_i(|\boldsymbol{v}_j(x_i)|)$, we suggest to set the threshold for $\boldsymbol{v}_j$ as $\epsilon_j = \max_i(|\boldsymbol{v}_j(x_i)|)/n$ (n as the sample size) to accommodate the perturbation.

6. Experimental Results on Simulations and USPS Data Set.

6.1. *Simulation: Gaussian Type Components.* In this simulation, we examine the effectiveness of the proposed DaSpec algorithm on datasets generated from Gaussian mixture models. Each data set (size of 400) is sampled from a mixture of six bivariate Gaussians, while the size of each group follows a Multinomial distribution ($n = 400$, and $p_1 = \cdots = p_6 = 1/6$). The mean and standard deviation of each Gaussian are randomly drawn from a Uniform on $(-5, 5)$ and a Uniform on $(0, 0.8)$ respectively. Four data sets generated from this distribution are plotted in the left column of Figure 5. It is clear that the groups may be highly unbalanced and overlap each other. Therefore, rather than trying to separate all six components, we expect good clustering algorithms to identify groups with reasonable separation between high density areas.

The DaSpec algorithm is applied with parameters ω and ϵ_j data-adaptively chosen by the criteria described in Section 5.2. Taking the number of groups identified by our Daspec algorithm, the commonly used k-means algorithm and the spectral clustering algorithms proposed in Ng, et al. [9] (using the same ω as the DaSpec) are also tested to serve as baselines for comparison. As a common practice with k-means algorithm, fifty random initializations are carried out and the final results are from the one that minimizes the optimization criterion $\sum_{i=1}^{n}(x_i - y_{k(i)})^2$, where x_i is assigned to group $k(i)$ and $y_k = \sum_{i=1}^{n} x_i I(k(i) = k) / \sum_{i=1}^{n} I(k(i) = k)$.

Shown in the second column of Figure 5, the proposed DaSpec algorithm (with data-dependent parameter choice) identifies the number of separable groups, isolates potential outliers and groups data accordingly. The results are similar to the k-means algorithm results (the third column) when the groups are balanced and their shapes are close to round. In those cases, the k-means algorithm is expected to work well given that the the data in each group are well represented by their averages. The last column shows the results of Ng's spectral clustering algorithm, which sometimes assign data to one group even when they are actually far away.

In summary, for this simulated example, we find that the proposed DaSpec algorithm with data-adaptively chosen parameters identifies the number of separable groups reasonably well and produces good clustering results when the separations are enough. It is also interesting to note that the algorithm also isolates possible "outliers" into a separate group, so they do not affect the clustering results on the majority data. The proposed algorithm also competes well against the commonly used k-means and spectral clustering algorithms.

6.2. *Simulation: Beyond Gaussian Components.* We now compare the performance of the aforementioned clustering algorithms on data sets that contain non-Gaussian groups, various levels of noise, and possible outliers. Data set $\mathcal{D}_1$ contains three well-separable groups and an outlier in $\mathbb{R}^2$. The first group of data are generated by adding independent Gaussian noise $N((0,0)^T, 0.15^2 I)$ to 200 uniform samples from three fourth of a ring with radius 3, which is from the same distribution as those plotted the right panel of Figure 3. The second group includes 100 data points sampled from a bivariate Gaussian $N((3,-3)^T, 0.5^2 I)$ and the last group has only 5 data points sampled from a bivariate Gaussian $N((0,0)^T, 0.3^2 I)$. Finally, one outlier is located at $(5,5)^T$. Given $\mathcal{D}_1$, three more data sets ($\mathcal{D}_2$, $\mathcal{D}_3$, and $\mathcal{D}_4$) are created by gradually adding independent Gaussian noise (with standard deviations 0.3, 0.6, 0.9 respectively). The scatter plots of the four datasets are shown in the left column of Figure 6. It is clear that the degree of separation decreases from top to bottom.

Similar to the previous simulation, we examine the DaSpec algorithm with data-adaptively chosen parameters, the k-means and Ng's spectral clustering algorithms on these data sets. the later two algorithms are tested under two different assumptions on the number of groups: the number (G) identified by the DaSpec algorithm or one group less ($G-1$). Note that the DaSpec algorithm claims only one group for $\mathcal{D}_4$, so the other two algorithms are skipped.

The DaSpec algorithm (the second column in the right panel of Figure 6) produces reasonable number of groups and clustering results. For the perfectly separable case in $\mathcal{D}_1$, three groups are identified and the one outlier is isolated out. It is worth to note that the incomplete ring is separated from other groups, which is not a simple task for algorithms based on group centroids. We also see that the DaSpec algorithm starts to combine inseparable groups as the components become less separable.

Not surprisingly, the k-means algorithms (the third and fourth columns) do not perform well because of the presence of the non-Gaussian component, unbalanced groups and outliers. Given enough separations, the spectral clustering algorithm reports reasonable results (the fifth and sixth columns). However, it is sensitive to outliers and the specification of the number of groups.

6.3. *Application: USPS Zip Code Data.*

In addition to the simulation studies, we use a high-dimensional U.S. Postal Service (USPS) digit data set to test the DaSpec algorithm and the rational behind it. The data set contains normalized handwritten digits, automatically scanned from envelopes by the USPS. The images here have been rescaled and size normalized, resulting in 16×16 grayscale images (see Le Cun, et al., [7] for details). Each image is treated as a vector in $\mathbb{R}^{256}$. In this experiment, 658 "3"s, 652 "4"s, and 556 "5"s in the training data are pooled together as our sample (size

1866).

To apply the DaSpec algorithm on this high-dimensional data set, we try to chose the kernel bandwidth either data-adaptively or manually by maximizing the accuracy of the clustering results compared to the known group labels. As we expected, the data-adaptively chosen bandwidth ($\omega = 0.82$) is too small and the algorithm claims more than three groups. By comparing the results to the know labels using different bandwidths, we chose $\omega = 2$ and use $\epsilon_j = \max(|\boldsymbol{v}_j|)/n$.

With this manually selected bandwidth, three eigenvectors $\boldsymbol{v}_1$, $\boldsymbol{v}_{16}$ and $\boldsymbol{v}_{49}$ are identified as no sign changes up to ϵ_j. To visualize the results, the digits are first ranked by an decreasing order of a given $|\boldsymbol{v}_j|$, $j = 1, \ldots, 50$, and the 1^{st}, 36^{th}, 71^{st}, $\cdots$, 316^{th} digits according to that order are shown in each row of Figure 7. Beside visualizing the images, we show the scatter plot of the data embedded in the top three eigenvectors in the left panel of Figure 8 and that of the 1^{st}, 16^{th} and 49^{th} eigenvectors in the middle panel.

The results strongly support our rationale of skipping certain top eigenvectors in our algorithm. As shown in Figure 7 and the right panel of Figure 8, the digits with large absolute values of the top three eigenvectors all represent number "4". Hence, the space spanned by the top three eigenvectors of Kn does not provide much information about "3" and "5", which is suggested by our theoretical analysis (Mixture property of top spectrum in Section 4.2). Actually, the digits with large absolute values of the top 15 eigenvectors all represent number "4", which lead to the failure of clustering algorithms only using the top eigenvectors of K_n. As shown in the right panel of Figure 8, the k-means algorithm based on top eigenvectors (normalized as suggested in Scott and Longuet-Higgins [14]) reports accuracies below 80% and it reaches the best performance as the 49^{th} eigenvector is included.

On the other hand, the three eigenvectors identified by our algorithm do present the three groups of digits "3", "4" and "5" nearly perfectly. As we also expected, the digits corresponding to large absolute values of an eigenvector are from the same group. The scatter plot of embedded data in the three identified eigenvectors shown in the right panel of Figure 8 perfectly agrees with what the theoretical results suggested. The overall accuracy of the DaSpec algorithm stands at 93.57%, the same as the k-means using only the 1st, 16th and 49th eigenvectors. Assuming three groups, k-means using the original input vector reports a 93.3% accuracy and Ng's spectral clustering with a manually selected Gaussian bandwidth ($\omega = 6.7$) stands at 92.93%, both works relatively well since the groups are balanced and the separations between groups are relatively large.

7. Conclusions and Discussion. Motivated by recent developments in kernel and spectral methods, we study the connection between a probability distribution and the associated convolution operator. For a convolution operator defined by a radial kernel with a fast tail decay, we show that the top eigenfunctions of the operator defined on a mixture distribution is approximately a combination of the top eigenfunctions of each component. The separation condition is mainly based on the overlap between high-density components, instead of their explicit parametric forms, and thus is quite general. These theoretical results explain why the top eigenvectors of kernel matrix may preserve the clustering information but not always do so. More importantly, our results reveal that not every component will contribute to the top few eigenfunctions of convolution operator $\mathcal{K}_p$ because the size and configuration of a component decide the corresponding eigenvalues. Hence the top eigenvectors of the kernel matrix may or may not preserve all clustering information, which explains some empirical observations of certain spectral clustering methods.

Following the theoretical analyses, a Data Spectroscopic clustering algorithm, DaSpec, is proposed, based on finding eigenvectors with no sign changes (not necessarily the top ones). Comparing to commonly used k-means and spectral clustering algorithms, DaSpec is very simple to implement, naturally provides an estimator of the number of separate groups, and handles the unbalancing weight and outliers well. More importantly, unlike k-means and certain spectral clustering algorithms, DaSpec does not require random initialization, which is a potentially significant advantage in practice. Simulations and an application to high-dimensional digit clustering show favorable results compared to k-means and spectral clustering algorithms. For practical applications, we also provide some guidelines for choosing the algorithm parameters.

Our analyses and discussions on connection to other spectral or kernel methods shed light on why radial kernels, such as a Gaussian kernel, perform well in many classification and clustering algorithms. We expect that this line of investigation would also prove fruitful in understanding other kernel algorithms, such as Support Vector Machines.

APPENDIX

Proof of Theorem 2: For a semi-positive definite kernel $K(x, y) > 0$, we first show the top eigenfunction ϕ_0 of $\mathcal{K}_p$ has no sign change on the support of the distribution. We define $R^+ = \{x \in \mathbb{R}^d : \phi_0(x) > 0\}$, $R^- = \{x \in \mathbb{R}^d : \phi_0(x) < 0\}$ and $\phi_0^*(x) = |\phi_0(x)|$. It is clear that $\int [\phi_0^*(x)]^2 p(x)dx = \int [\phi_0(x)]^2 p(x)dx$.

Assume that $P(R^+) > 0$ and $P(R^-) > 0$, we will show that

$$\int\int K(x,y)\phi_0^*(x)\phi_0^*(y)p(x)p(y)dxdy > \int\int K(x,y)\phi_0(x)\phi_0(y)p(x)p(y)dxdy, \tag{A.1}$$

which contradicts with the assumption that $\phi_0(\cdot)$ is the eigenfunction associated with the largest eigenvalue. Denoting $g(x,y) = K(x,y)\phi_0(x)\phi_0(y)p(x)p(y)$ and $g^*(x,y) = K(x,y)\phi_0^*(x)\phi_0^*(y)p(x)p(y)$, we have

$$\int_{R^+(x)}\int_{R^+(y)} g^*(x,y)dxdy = \int_{R^+(x)}\int_{R^+(y)} g(x,y)dxdy,$$

and the equation also holds on region $R^-(x) \times R^-(y)$. However, over the region $\{(x,y) : x \in R^+ \text{ and } y \in R^-\}$, we have

$$\int_{R^+(x)}\int_{R^-(y)} g^*(x,y)dxdy > \int_{R^+(x)}\int_{R^-(y)} g(x,y)dxdy,$$

since $K(x,y) > 0$, $\phi_0(x) > 0$, and $\phi_0(y) < 0$. The inequality holds on $\{(x,y) : x \in R^- \text{ and } y \in R^+\}$. Putting four integration regions together, we arrive inequality (A.1). Therefore, the assumptions $P(R^+) > 0$ and $P(R^-) > 0$ can not be true at the same time, which implies that $\phi_0(\cdot)$ has no sign changes on the support of the distribution.

Now consider $\forall x \in \mathbb{R}^d$. we have

$$\lambda_0\phi_0(x) = \int K(x,y)\phi_0(y)p(y)dy.$$

Given the facts that $\lambda_0 > 0$, $K(x,y) > 0$, and $\phi_0(y)$ have the same sign on the support, it is straightforward to see that $\phi_0(x)$ has sign changes and has full support in $\mathbb{R}^d$. Finally, the isolation of (λ_0, ϕ_0) follows. If there exist another ϕ that shares the same eigenvalue λ_0 with ϕ_0, they both have no sign change and have full support on $\mathbb{R}^d$. Therefore $\int \phi_0(x)\phi(x)p(x)dx > 0$ and it contradicts with the orthogonality between eigenfunctions. □

Proof of Theorem 3: By definition, the top eigenvalue of $\mathcal{K}_p$ satisfies:

$$\begin{aligned}\lambda_0 &= \max_{f:\int f^2dP=1}\iint K(x,y)f(x)f(y)p(x)p(y)dxdy \\ &= \max_f \frac{\iint K(x,y)f(x)f(y)p(x)p(y)dxdy}{\int[f(x)]^2p(x)dx}.\end{aligned}$$

For any function f,

$$\begin{aligned}
& \iint K(x,y)f(x)f(y)p(x)p(y)dxdy \\
= \; & (\pi^1)^2 \iint K(x,y)f(x)f(y)p^1(x)p^1(y)dxdy \\
& + (\pi^2)^2 \iint K(x,y)f(x)f(y)p^2(x)p^2(y)dxdy \\
& + 2\,\pi^1\pi^2 \iint K(x,y)f(x)f(y)p^1(x)p^2(y)dxdy \\
\leq \; & (\pi^1)^2\lambda_0^1 \int [f(x)]^2 p^1(x)dx + (\pi^2)^2\lambda_0^2 \int [f(x)]^2 p^2(x)dx \\
& + 2\,\pi^1\pi^2 \iint K(x,y)f(x)f(y)p^1(x)p^2(y)dxdy
\end{aligned}$$

Now we concentrate on the last term:

$$\begin{aligned}
& 2\pi^1\pi^2 \iint K(x,y)f(x)f(y)p^1(x)p^2(y)dxdy \\
= \; & 2\pi^1\pi^2 \iint \left[K(x,y)[p^1(x)]^{1/2}[p^2(y)]^{1/2}\right]\left[f(x)f(y)[p^1(x)]^{1/2}[p^2(y)]^{1/2}\right] dxdy \\
\leq \; & 2\pi^1\pi^2 \sqrt{\iint [K(x,y)]^2 p^1(x)p^2(y)dxdy}\sqrt{\iint [f(x)]^2[f(y)]^2 p^1(x)p^2(y)dxdy} \\
\leq \; & 2\pi^1\pi^2 \sqrt{\iint [K(x,y)]^2 p^1(x)p^2(y)dxdy}\sqrt{\iint [f(x)]^2[f(y)]^2 p^1(x)p^2(y)dxdy} \\
= \; & 2\sqrt{\pi^1\pi^2 \iint [K(x,y)]^2 p^1(x)p^2(y)dxdy}\sqrt{\pi^1 \int [f(x)]^2 p^1(x)dx}\sqrt{\pi^2 \int [f(y)]^2 p^2(y)dy} \\
\leq \; & \sqrt{\pi^1\pi^2 \iint [K(x,y)]^2 p^1(x)p^2(y)dxdy}\left(\int [f(x)]^2\pi^1 p^1(x)dx + \int [f(x)]^2 \pi^2 p^2(x)dx\right) \\
= \; & r \int [f(x)]^2 p(x)dx
\end{aligned}$$

where $r = (\pi^1\pi^2 \iint [K(x,y)]^2 p^1(x)p^2(y)dxdy)^{1/2}$. Thus,

$$\begin{aligned}
\lambda_0 \; = \; & \max_{f:\int [f(x)]^2 p(x)dx=1} \iint K(x,y)f(x)f(y)p(x)p(y)dxdy \\
\leq \; & \max_{f:\int [f(x)]^2 p(x)dx=1} \left[\pi^1\lambda_0^1 \int [f(x)]^2 \pi^1 p^1(x)dx + \pi^2\lambda_0^2 \int [f(x)]^2 \pi^2 p^2(x)dx + r\right] \\
\leq \; & \max(\pi^1\lambda_0^1, \pi^2\lambda_0^2) + r
\end{aligned}$$

The other side of the equality is easier to prove. Assuming $\pi^1\lambda_0^1 > \pi^2\lambda_0^2$ and taking the top eigenfunction ϕ_0^1 of $\mathcal{K}_{p^1}$ as f, we derive the following results by using the same decomposition on $\iint K(x,y)\phi_0^1(x)\phi_0^1(y)p(x)p(y)dxdy$

and the facts that $\int K(x,y)\phi_0^1(x)p^1(x)dx = \lambda_0^1\phi_0^1(y)$ and $\int[\phi_0^1(x)]^2p^1(x)dx = 1$.

$$\begin{aligned}
\lambda_0 &\geq \frac{\iint K(x,y)\phi_0^1(x)\phi_0^1(y)p(x)p(y)dxdy}{\int[\phi_0^1(x)]^2p(x)dx} \\
&= \frac{(\pi^1)^2\lambda_0^1 + (\pi^2)^2\iint K(x,y)\phi_0^1(x)\phi_0^1(y)p^2(x)p^2(y)dxdy + 2\pi^1\pi^2\lambda_0^1\int[\phi_0^1(x)]^2p^2(x)dx}{\pi^1 + \pi^2\int[\phi_0^1(x)]^2p^2(x)dx} \\
&= \pi^1\lambda_0^1\left(\frac{\pi^1 + 2\pi^2\int[\phi_0^1(x)]^2p^2(x)dx}{\pi^1 + \pi^2\int[\phi_0^1(x)]^2p^2(x)dx}\right) + \frac{(\pi^2)^2\iint K(x,y)\phi_0^1(x)\phi_0^1(y)p^2(x)p^2(y)dxdy}{\pi^1 + \pi^2\int[\phi_0^1(x)]^2p^2(x)dx} \\
&\geq \pi\lambda_0^1.
\end{aligned}$$

This completes the proof. □

ACKNOWLEDGEMENT

The authors would like to thank Yoonkyung Lee, Prem Goel, Joseph Verducci, and Donghui Yan for helpful discussions, suggestions and comments.

REFERENCES

[1] M. Belkin and P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Computation, 15(6) (2003), pp. 1373C1396.

[2] F. Chung, *Spectral graph theory*, vol. 92 of CBMS Regional Conference Series in Mathematics, Washington, 1997.

[3] I. Dhillon, Y. Guan, and B. Kulis, *A unified view of kernel k-means, spectral clustering, and graph partitioning*, Tech. Rep. UTCS TF-04-25, University of Texas at Austin, 2005.

[4] P. Diaconis, S. Goel, and S. Holmes, *Horseshoes in multidimensional scaling and kernel methods*, To appear in Annals of Applied Statistics, (2007).

[5] L. Hagen and A. Kahng, *New spectral methods for ratio cut partitioning and clustering*, IEEE Trans. Computer-Aided Design, 11(9) (1992), pp. 1074–1085.

[6] V. Koltchinskii and E. Giné, *Random matrix approximation of spectra of integral operators*, Bernoulli, 6 (2000), pp. 113 – 167.

[7] Y. Le Cun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, and L. Jackel, *Handwritten digit recognition with a backpropogation network*, in Advances in Neural Information Processing Systems, D. Touretzky, ed., vol. 2, Morgan Kaufman, Denver CO, 1990.

[8] M. Melia and J. Shi, *A random walks view of spectral segmentation*, in AI and STATISTICS (AISTATS), 2001.

[9] A. Ng, M. Jordan, and Y. Weiss, *On spectral clustering: Analysis and an algorithm*, in Advances in Neural Information Processing Systems 14, T. Dietterich, S. Becker, and Z. Ghahramani, eds., MIT Press, 2002, pp. 955 – 962.

[10] B. N. Parlett, *The summetric Eigenvalue Problem*, Prentice Hall, 1980.

[11] P. Perona and W. Freeman, *A factorization approach to grouping*, in Proceedings of ECCV, 1998, pp. 655–670.

[12] B. Schölkopf and A. Smola, *Learning with Kernels*, MIT Press, Cambridge, MA, 2002.

[13] B. Schölkopf, A. Smola, and K. R. Müller, *Nonlinear component analysis as a kernel eigenvalue problem*, Neural Computation, 10 (1998), pp. 1299–1319.

[14] G. Scott and H. Longuet-Higgins, *Feature grouping by relocalisation of eigenvectors of proxmity matrix*, in Proceedings of British Machine Vision Conference, 1990.

[15] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence, 22 (2000), pp. 888–905.

[16] T. Shi, M. Belkin, and B. Yu, *Data spectroscopy: Learning mixture models using eigenspaces of convolution operators*, in Proceedings of the 25^{th} International Conference on Machine Learning, 2008.

[17] V. Vapnik, *The Nature of Statistical Learning*, Springer, 1995.

[18] U. von Luxburg, *A turorial on spectral clustering*, Statistics and Computing, 17(4) (2007), pp. 395 – 416.

[19] Y. Weiss, *Segmentation using eigenvectors: A unifying view*, in Proceedings of the International Conference on Computer Vision, 1999, pp. 975–982.

[20] C. K. Williams and M. Seeger, *The effect of the input density distribution on kernel-based classifiers*, in Proceedings of the 17th International Conference on Machine Learning, P. Langley, ed., San Francisco, California, 2000, Morgan Kaufmann, pp. 1159–1166.

[21] H. Zhu, C. Williams, R. Rohwer, and M. Morcinie, *Gaussian regression and optimal finite dimensional linear models*, in Neural networks and machine learning, C. Bishop, ed., Berlin: Springer-Verlag, 1998.

Tao Shi
Department of Statistics
The Ohio State University
1958 Neil Avenue, Cockins Hall 404
Columbus, OH 43210-1247
E-mail: taoshi@stat.osu.edu

Mikhail Belkin
Department of Computer Science and Engineering
The Ohio State University
2015 Neil Avenue, Dreese Labs 597
Columbus, OH 43210-1277
E-mail: mbelkin@sce.osu.edu

Bin Yu
Department of Statistics
University of California, Berkeley
367 Evans Hall
Berkeley, CA 94720-3860
E-mail: binyu@stat.berkeley.edu

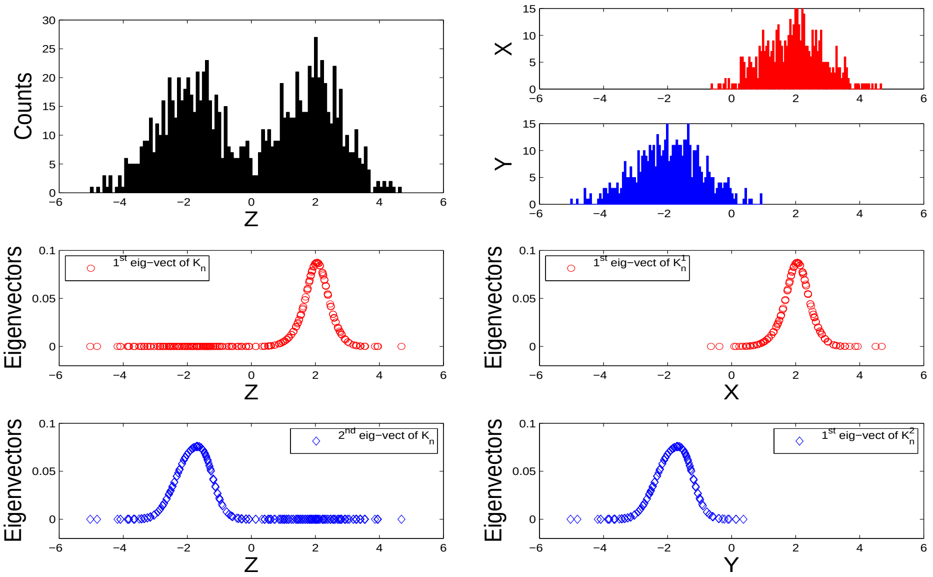

FIG 1. *Eigenvectors of a Gaussian kernel matrix ($\omega = 0.3$) of 1000 data sampled from a Mixture Gaussian distribution $0.5N(2, 1^2) + 0.5N(-2, 1^2)$. Left panels: Histogram of the data (top), first eigenvector of K_n (middle), and second eigenvector of K_n (bottom). Right panels: Histograms of data from each component (top), first eigenvector of K_n^1 (middle), and first eigenvector of K_n^2 (bottom).*

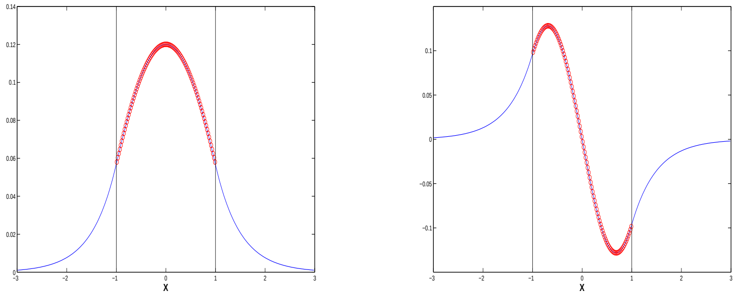

FIG 2. *Top two eigenfunctions of the exponential kernel with bandwidth $\omega = 0.5$ and the uniform distribution on $[-1, 1]$.*

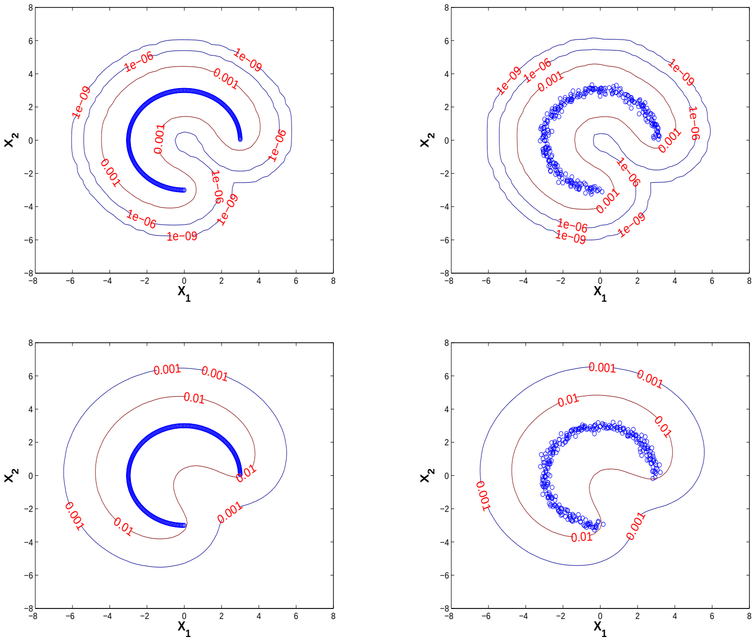

Fig 3. *Contours of the top eigenfunction of $\mathcal{K}_p$ for Gaussian (upper panels) and exponential kernels (lower panels) with bandwidth* 0.7*. The curve is 3/4 of a ring with radius 3 and independent noise of standard deviation 0.15 added in the right panels.*

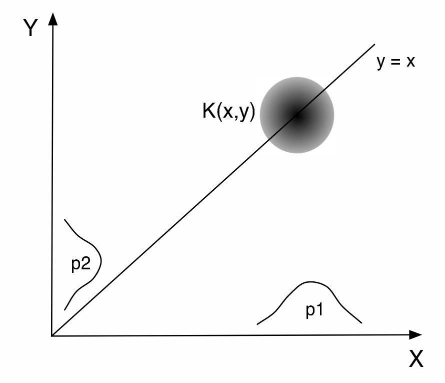

Fig 4. *Illustration of separation condition (4.1) in Theorem 3.*

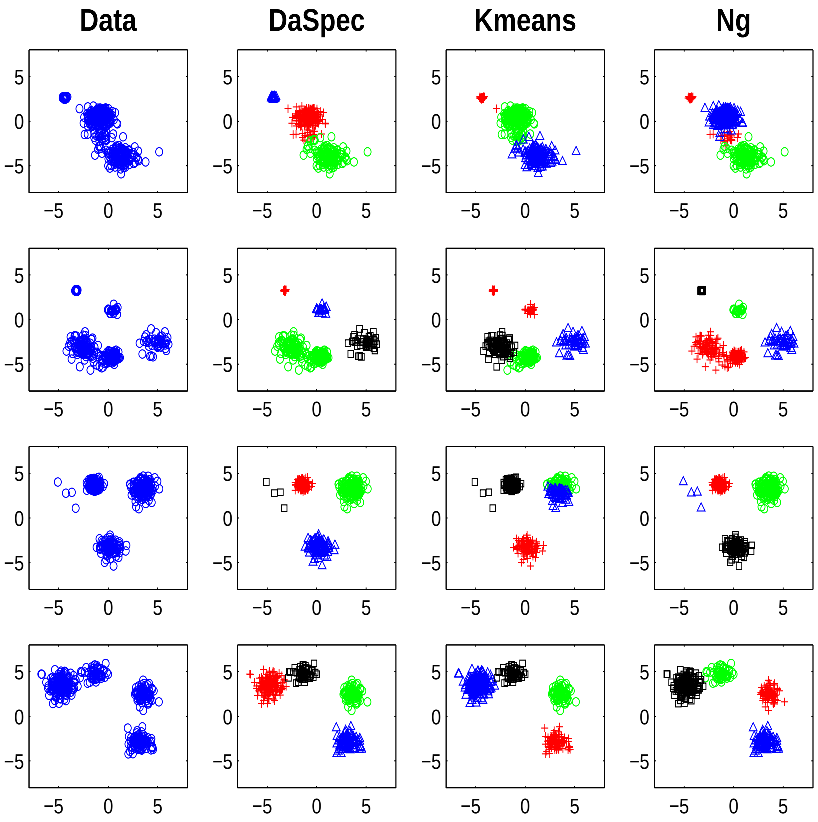

FIG 5. *Clustering results on four simulated data sets described in Section 6.1. First column: scatter plots of data; Second column: results the proposed spectroscopic clustering algorithm; Third column: results of the k-means algorithm; Fourth column: results of the spectral clustering algorithm (Ng, et al. [9]).*

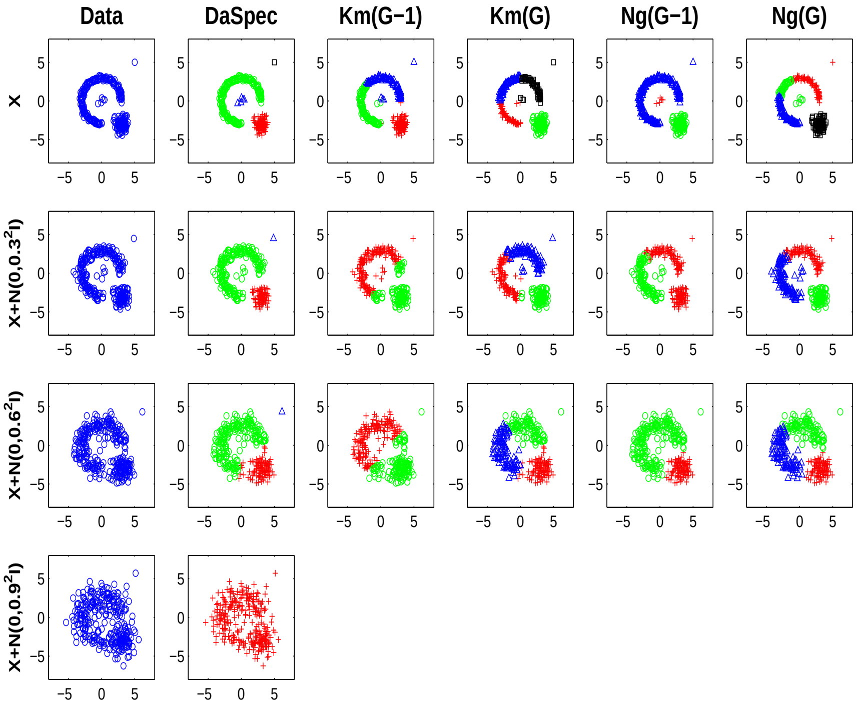

FIG 6. *Clustering results on four simulated data sets described in Section 6.2. First column: scatter plots of data; Second column: labels of the G identified groups by the proposed spectroscopic clustering algorithm; Third and forth columns: k-means algorithm assuming $G-1$ and G groups respectively; Fifth and sixth columns: spectral clustering algorithm (Ng et al. [9]) assuming $G-1$ and G groups respectively.*

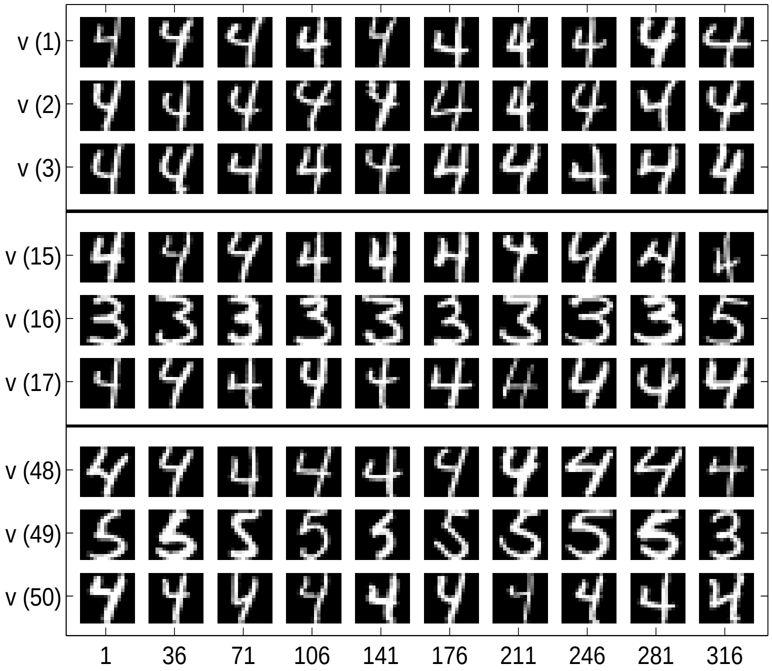

FIG 7. *Digits ranked by the absolute value of eigenvectors $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, ..., $\boldsymbol{v}_{50}$. The digits in each row correspond to the 1^{st}, 36^{th}, 71^{st}, $\cdots$, 316^{th} largest absolute value of the selected eigenvector. Three eigenvectors, $\boldsymbol{v}_1$, $\boldsymbol{v}_{16}$, and, $\boldsymbol{v}_{49}$, are identified by our DaSpec algorithm.*

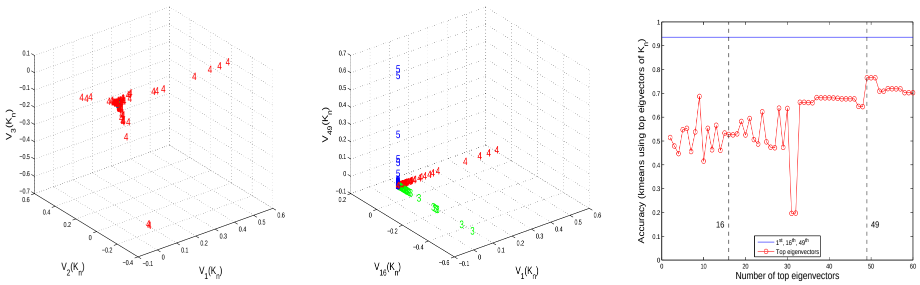

FIG 8. *Left: Scatter plots of digits embedded in the top three eigenvectors; Middle: Digits embedded in the 1^{st}, 16^{th} and 49^{th} eigenvectors; Right: Accuracy of kmeans algorithms using different number of top eigenvectors of K_n*