SIGN-BASED SPECTRAL CLUSTERING

H. T. Kung *Dario Vlah*
{*htk, dario*}*@eecs.harvard.edu*

Harvard School of Engineering and Applied Sciences
Cambridge, MA 02138

Abstract

Sign-based spectral clustering performs data grouping based on signs of components in the eigenvectors of the input. This paper introduces the concept of sign-based clustering, proves some of its basic properties and describes its use in applications. It is shown that for certain applications where a relatively small number of clusters are sought the sign-based approach can greatly simplify clustering by just examining the signs of components in the eigenvectors, while improving the speed and robustness of the clustering process. For other such applications, it can provide useful initial approximations in improving the performance of cluster searching heuristics such as k-means.

1. INTRODUCTION

In this paper, we consider the problem of clustering sensor nodes in a sensor-target scenario, where there are N sensor nodes taking measurements of M targets. We will classify the sensors in terms of their measurements of the targets. That is, given a $M \times N$ input matrix $\mathbf{A}$ with entries being the sensor-to-target measurements, we are interested in clustering the N sensor nodes based on $\mathbf{A}$.

Spectral clustering refers to a type of methods which classify nodes based on the eigenstructure of the $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ matrix which represents some inter-node relationship. By exploiting the fact that the eigenstructure conveniently captures important properties of the input data, spectral clustering has a variety of applications in areas such as communications, sensing, signal processing, information retrieval, security and networking [1–6]. However, when applying spectral clustering to real-world data, we often face the difficult task of choosing proper heuristics in searching for clusters in the spectral space. These heuristics may involve decisions such as choosing the proper number of clusters to find, setting various clustering thresholds and obtaining initial approximations.

In this paper, we take a fresh view in approaching spectral clustering. We note that clustering becomes difficult often due to the high complexity in the heuristics used. Because these heuristics are designed to handle all sorts of data, they may not be among the most effective methods for quantized data which inherently lead to more separation for clustering purposes. For example, when the k-means based spectral clustering [7] is used, it is essential to have a correct initial setting on the number of clusters, otherwise the method could incur a large computing cost or may converge very slowly [8]. In many practical applications, input data assume a relatively small set of discrete values for reasons such as signal noise and inaccuracy of instruments, and a relatively small number of clusters are sought. This allows clustering to take place at a coarse level. In these situations it would make sense to explore the use of more direct clustering algorithms such as the sign-based spectral approach of this paper rather than relying on less deterministic search-based heuristics.

We show that for some of these quantized problems, we can greatly simplify spectral clustering while improving its speed and robustness. In particular, we can cluster data by just examining the signs of components in the eigenvectors of the input data.

2. FIRST EXAMPLE: A SENSOR-TARGET SCENARIO

To illustrate sign-based clustering, we consider a simple sensor-target example. There are 19 sensors and 3 targets placed on a line, numbered 0-18 and 0-2, respectively. The line has 91 uniformly spaced grid points, labeled as 0, 1, ..., 90. Sensor i and target j are located at grid points $5i$ and $45j$, respectively. We model the measurement of target j obtained by sensor i as a function of the difference between their location values. We discretize the measurement by using a step function shown in Figure 1. For example, sensor 3 is at location 15 and target 1 is at location 45; thus, their locations differ by 30. This means that according to the quantized measurement function in Figure 1, sensor 3's measurement value at target 1 is 0. Note the

This material is based on research sponsored by Air Force Research Laboratory under agreement number FA8750-08-1-0191. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory or the U.S. Government.

978-1-4244-5711-3/10/$26.00 ©2010 IEEE

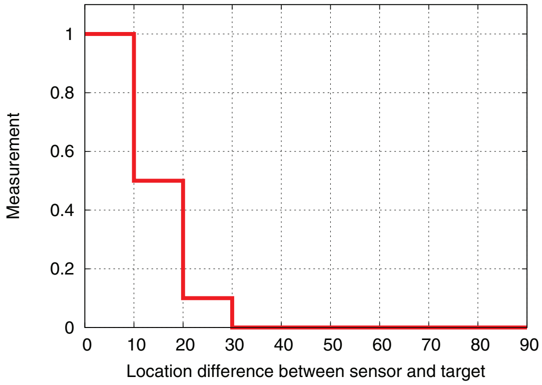


Fig. 1: Quantization step function associated with sensor measurements for the sensor-target example.

$$\mathbf{A}=\left[\begin{array}{ccccccccccccccccccc}1 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{10} & \frac{1}{10} & & & & & & & & & & & & & \\ & & & & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} & \frac{1}{2} & 1 & 1 & 1 & \frac{1}{2} & \frac{1}{2} & \frac{1}{10} & \frac{1}{10} & & & & \\ & & & & & & & & & & & & & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} & \frac{1}{2} & 1 & 1\end{array}\right]$$

Fig. 2: Quantized 3×19 sensor-to-target matrix $\mathbf{A}$ based on the step function of Figure 1. The empty entries are zeros.

tail of the step function is truncated at distance equal to 30. Based on the quantized measurement values we can define a 3×19 sensor-to-target matrix $\mathbf{A}$ in Figure 2. Visually we can see three clusters of sensors each corresponding to one of the three targets, as depicted in Figure 3.

We demonstrate that sign-based spectral clustering can discover these three sensor clusters. To this end, we first form $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, which is a 19×19 sensor-to-sensor matrix. We can check that $rank(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = 3$. Thus $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ has three eigenvectors. Figure 4 depicts the components of these three eigenvectors. We define a sign sequence for a set of eigenvectors for a component position to be composed of component signs of these eigenvectors at this component position. We note that the components of the three eigenvectors in Figure 4 exhibit sign sequences $(+,-,+)$, $(+,-,-)$ and $(+,+,+)$ for sensors 0 through 5, 6-12 and 13 through 18, respectively. Thus, the three clusters of sensors identified by the three sign sequences indeed correspond to those depicted in Figure 3.

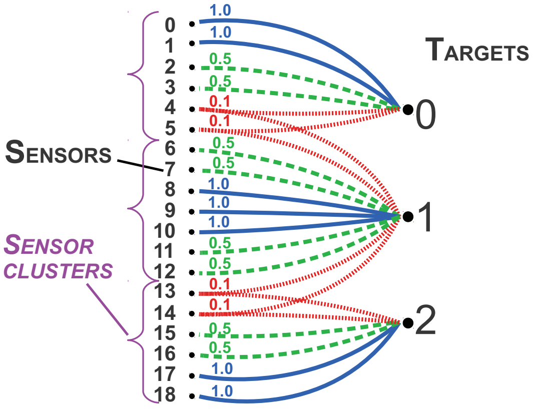
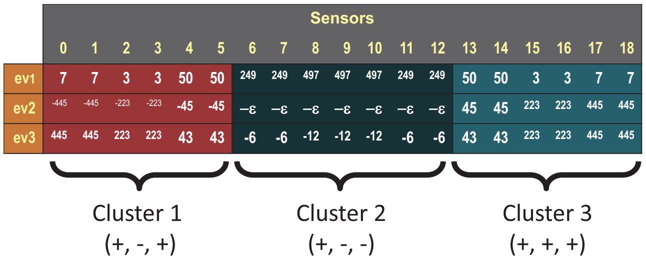


Fig. 3: Three sensor clusters—sensors 0 through 5, 6–12, and 13 through 18—corresponding to the three targets.



	Sensors																		
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
ev1	7	7	3	3	50	50	249	249	497	497	497	249	249	50	50	3	3	7	7
ev2	-445	-445	-223	-223	-45	-45	$-\varepsilon$	$-\varepsilon$	$-\varepsilon$	$-\varepsilon$	$-\varepsilon$	$-\varepsilon$	$-\varepsilon$	45	45	223	223	445	445
ev3	445	445	223	223	43	43	-6	-6	-12	-12	-12	-6	-6	43	43	223	223	445	445

Cluster 1
(+, -, +)
Cluster 2
(+, -, -)
Cluster 3
(+, +, +)

Fig. 4: Eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ for $\mathbf{A}$ given in Figure 2 and sign-based clustering of sensors. The first three eigenvectors are shown as ev1, ev2 and ev3. The symbol $-\varepsilon$ stands for a small negative value.

3. SECOND EXAMPLE: LINE DETECTION VIA HOUGH TRANSFORM

Here we give another simple illustration concerning detection of lines based on the Hough transform [9]. As depicted in Figure 5, we are given 14 points, P_i's, lying on five lines, Line j's. In the sensor-target terminology of this paper, points and lines here assume the role of "sensors" and "targets," respectively. More precisely, each point will "vote" on every line as follows: vote 10 if the point is on the line and the line intersects another line, 5 if the point is on the line and the line does not intersect any other line, 2 if the point is on an intersecting line, and 1 otherwise. Figure 6 shows the 5×14 point-to-line matrix $\mathbf{A}$ capturing the votes. Note that votes are discrete in the sense that they assume only a few values. This is analogous to the discrete measurements imposed by the quantization step function in the preceding sensor-target example.

We demonstrate that sign-based spectral clustering can discover clusters of points in terms of their relationships to the given lines. We first form $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, which is a 14×14 point-to-point matrix. We can check that $rank(\mathbf{A}^{\mathrm{T}}\mathbf{A}) = 5$. We then examine the five eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ shown in Figure 7. We note that the components of eigenvectors exhibit sign sequences (+, +, - , 0, 0), (+, 0, -, 0, 0), (+, -, -, 0, 0), (+, 0, +, +, -), (+, 0, +, -, -) and (+, 0, +, +, +) for points $\{P_0, P_3\}$, $\{P_2\}$, $\{P_4, P_5\}$, $\{P_6, P_7, P_8\}$, $\{P_9, P_{10}, P_{11}\}$ and $\{P_{12}, P_{13}, P_{14}\}$, respectively. One can check that these clusters of points identified by the sign sequences indeed correspond to the 5 given lines, and the intersecting point of Line 1 and Line 2. Further, Figure 7 shows that as we increase the

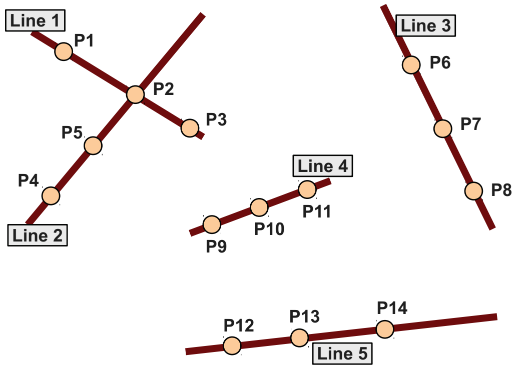
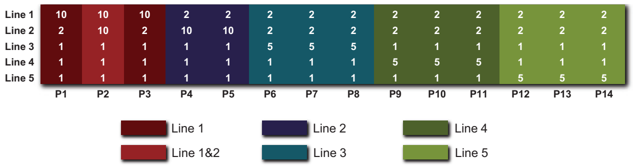


Fig. 5: Line detection via Hough transform in a scenario with 14 points and 5 lines.



	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
Line 1	10	10	10	2	2	2	2	2	2	2	2	2	2	2
Line 2	2	10	2	10	10	2	2	2	2	2	2	2	2	2
Line 3	1	1	1	1	1	5	5	5	1	1	1	1	1	1
Line 4	1	1	1	1	1	1	1	1	5	5	5	1	1	1
Line 5	1	1	1	1	1	1	1	1	1	1	1	5	5	5

Line 1
Line 2
Line 4
Line 1&2
Line 3
Line 5

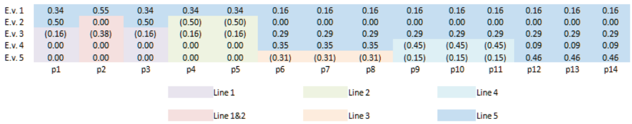
Fig. 6: Discrete 5×14 point-to-line matrix $\mathbf{A}$ resulting from the voting of each of the 14 points on the 5 given lines.

number of eigenvectors from 2 to 5 the sign sequences will reveal a finer clustering structure.

4. SIGN-BASED SPECTRAL CLUSTERING

In this section, we show that the illustrative results in the preceding two sections are mathematical consequences, and not merely coincidences.

We assume the following throughout the paper. We are given a set of N sensor nodes and M targets. Let $\mathbf{A}$ be the $M \times N$ input matrix which captures sensor measurements of targets. We assume that measurements are non-negative real numbers. Thus $\mathbf{A}$ is non-negative, and so is $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. In addition, we assume that $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is irreducible. This is not limiting in view of our goals of clustering sensors, for two groups of sensors which do not share any target of which they have positive measurements already belong to two clusters by definition. We assume quantized input, that is, discrete measurement values of sensor nodes on targets (i.e., entries of $\mathbf{A}$) assume only a relatively small number of values. In the two examples described above, the $\mathbf{A}$ matrix corresponds to that in Figure 2 or 6.



	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14
E.v. 1	0.34	0.55	0.34	0.34	0.34	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16	0.16
E.v. 2	0.50	0.00	0.50	(0.50)	(0.50)	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E.v. 3	(0.16)	(0.38)	(0.16)	(0.16)	(0.16)	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29	0.29
E.v. 4	0.00	0.00	0.00	0.00	0.00	0.35	0.35	0.35	(0.45)	(0.45)	(0.45)	0.09	0.09	0.09
E.v. 5	0.00	0.00	0.00	0.00	0.00	(0.31)	(0.31)	(0.31)	(0.15)	(0.15)	(0.15)	0.46	0.46	0.46

Line 1
Line 2
Line 4
Line 1&2
Line 3
Line 5

Fig. 7: The five eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ for $\mathbf{A}$ given in Figure 6 and sign-based clustering of points. Negative numbers are shown in red and in parentheses. Sign sequences of successive eigenvectors identify increasingly refined clustering structures. At each eigenvector, clusters identified thus far are denoted in different colors.

We partition nodes into *equivalence sets* by including in the same set those nodes which assume the same measurement value for each of the M targets. Thus for the example of Figure 3, the equivalence sets are sensors 0-1, 2-3, 4-5, 6-7, 8-10, 11-12, 13-14, 15-16 and 17-18. In general, if entries of $\mathbf{A}$ can assume any of a given set of α values, then the number of equivalence sets is bounded by $\min(N, \alpha^M)$. For example, for $\mathbf{A}$ in Figure 2, $N = 19$, $M = 3$ and $\alpha = 4$.

We formally define what we mean by clusters. We give two definitions of clusters based on values or signs of components in the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ and describe their relationship.

4.1. Value-based Clustering

A subset of nodes is said to form a *value-based cluster* if, for each eigenvector of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, its components corresponding to these nodes assume the same value. We note the following Theorem:

Theorem 1: Nodes are in an equivalence set if and only if they belong to the same value-based cluster.

Proof: Consider any two nodes in an equivalence set. Suppose that the two nodes correspond to columns i and j of the input matrix A. Then the two columns must be identical. Next, consider $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. Since rows i and j of $\mathbf{A}^{\mathrm{T}}$ are identical, so are rows i and j of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. This means that if v is an eigenvector of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ corresponding to a non-zero eigenvalue λ, then components i and j of $(\mathbf{A}^{\mathrm{T}}\mathbf{A})v$ are identical. Since $(\mathbf{A}^{\mathrm{T}}\mathbf{A})v = \lambda v$, components i and j of v are identical. This means that the two given nodes must belong to the same value-based cluster.

Conversely, consider any two nodes belonging to the same value-based cluster. Suppose that the two nodes correspond to component positions i and j in the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. By expanding the singular value decomposition of $\mathbf{A}$ as a sum of outer products, we have:

$$\mathbf{A} = \sum_{h=1}^{N} \sigma_h \mathbf{u}_h \mathbf{v}_h^T \tag{1}$$

where σ_h are singular values and $\mathbf{u}_h$ and $\mathbf{v}_h$ are corresponding eigenvectors of $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ and $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, respectively. Since, for each h, $\mathbf{v}_h$ has the same component value at positions i and j, columns i and j of $\mathbf{u}_h\mathbf{v}_h^T$ are identical. Thus, by the outer product expression above, columns i and j of $\mathbf{A}$ are identical. This means that the two nodes are in the same equivalence set. ■

By Theorem 1, we note that those nodes which have the same measurement value for each target belong to the same value-based cluster. This provides a motivation for our definition of value-based clusters. The definition is natural in the sense it meets our expectation that nodes which have the same, or more generally, similar measurements should belong to the same cluster. For instance, in the sensor-target example of Figure 3, those sensors which have the same quantized measurements, such as sensors 8, 9 and 10 (see Figure 2), belong to the same value-based cluster. Similarly, in the line detection example of Figure 5, because points P4 and P5 have the same quantized measurements (see Figure 6), they belong to the same value-based cluster.

4.2. Sign-based Clustering

We now give the second definition of clusters. A subset of nodes is said to form a *sign-based cluster* if, for each eigenvector of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, the components corresponding to these nodes assume the same signum. That is, for any given eigenvector, its components corresponding to these nodes are either all positive or negative or zero.

The theorem below states that by examining component signs of any K eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, for any $1 < K \leq rank(\mathbf{A}^{\mathrm{T}}\mathbf{A})$, it is guaranteed that we can identify at least K sign-based clusters. This means that we will not need to look at many eigenvectors if we are only interested in finding a few sign-based clusters. Throughout the argument, we rely on a simple fact that eigenvectors belonging to different eigenvalues are orthogonal.

Theorem 2: Consider any K eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, for $1 < K \leq rank(\mathbf{A}^{\mathrm{T}}\mathbf{A})$. These eigenvectors will exhibit at least K distinct sign sequences in their components, thereby being able to identify at least K sign-based clusters.

Proof: For notational convenience, two vectors are said to be *sign-identical* if they have the same sign for each component position. For example, if one of the vectors has a positive value for a component position, then the other must also have a positive value for that component position. Furthermore, two such vectors are said to be *sign-complementary* if they have opposite signs for each component position. For example, if one of the vectors has a positive value for a component position, then the other must have a negative value for that component position. We note that sign-identical or sign-complementary vectors can not be orthogonal since their inner product is positive or negative, respectively, rather than zero.

Note that since we assume $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ is nonnegative and irreducible, by the Perron-Frobenius theorem the principal eigenvector corresponding to the largest eigenvalue must have all of its components being of the same sign [10], that is, they are either all positive or all negative. Any other eigenvector must have components with different signs, due to its orthogonality to the principal eigenvector. Thus, every non-principal eigenvector is associated with a *positive* and also a *negative* group of nodes which correspond to positive and negative components of the eigenvector. In contrast, the principal eigenvector is associated with either a positive or negative group, not both. We first consider the case where these non-principal eigenvectors have no zero components.

We will examine one by one the K given eigenvectors. The first eigenvector we examine is the principal eigenvector if it is present in the given set of K eigenvectors, otherwise the first eigenvector is any non-principal eigenvector in the set.

Consider any other eigenvector in the given set, which we call the second eigenvector here. Being orthogonal to each other, the first and second eigenvectors can not be sign-identical nor sign-complememtary. Thus the positive and negative groups of the second eigenvector must break at least one group of the first eigenvector into two groups of nodes, which are distinguishable by the sign sequences of the first and second eigenvectors.

We consider yet another eigenvector in the given set, called the third eigenvector. Being orthogonal to each other, the second and third eigenvectors can not be sign-identical nor sign-complementary. This means that the positive and negative groups of the third eigenvector must break at least one group associated with the second eigenvector. This results in at least three groups of nodes which are distinguishable by the sign sequences of the first, second and third eigenvectors.

This procedure can continue until all the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ have been considered. With K eigenvectors considered, we can identify at least K sign-based clusters.

We now consider the case where these K given eigenvectors may have zero components. Suppose that the second eigenvector has some zero components. Then for the other component positions, at least one of the first and second eigenvectors must have both positive and negative components, due to its orthogonality to the first eigenvector. This means that the first and second eigenvectors can already distinguish at least three sign-based clusters rather than two. Consider any other eigenvector in the given set and call it the third eigenvector. Being orthogonal to each other, the second and third eigenvectors can not be sign-identical nor sign-complementary on those component positions for which the second eigenvector has non-zero values. This means that either the third eigenvector has the same sign for these component positions or its groups break at least one group of the second eigenvector. For the latter case, the third eigenvector has identified at least one additional sign-based cluster and we are done. For the former case, the third eigenvector must assume the opposite sign for either (1) part or (2) all remaining of component positions. For case (1), the third eigenvector has identified an additional sign-based cluster and we are done. In addition, the third eigenvector has at least one fewer zero in its components than the second eigenvector. (So all the zeros will be chased out eventually.) For case (2), the third eigenvector has no zero components, so the next eigenvector must identify a new sign-based cluster. Although in this case the third eigenvector itself does not identify an additional sign-based cluster, this is okay since as noted earlier the first and second eigenvectors have already identified at least three sign-based clusters. This argument applies to any newly added eigenvector in the procedure which has zero components, beyond just the second eigenvector. ■

To illustrate the procedure used in the proof, consider a case with $N = 9$ nodes and $K = 4$ eigenvectors, with the first eigenvector being the principal eigenvector. Suppose that there are three 2-node sign-based clusters and one 3-node sign-based cluster. Then the four eigenvectors could have the following four distinct sign sequences: (+,+,-,-), (+,-,-,+), (+,0,-,+) and (+,0,+,+). These sign sequences, which are capable of distinguishing the four sign-based clusters, may be arranged as follows:

First e.v.:	+	+	+	+	+	+	+	+	+
Second e.v.:	+	+	-	-	-	0	0	0	0
Third e.v.:	-	-	-	-	-	-	-	+	+
Fourth e.v.:	-	-	+	+	+	+	+	+	+

The table in Figure 7 provides another illustration of the procedure used in the proof. We can see from the table that when more eigenvectors are used, their sign sequences will reveal finer sign-based clusters.

4.3. Relationships between Sign-based and Value-based Clusters

Here we describe some relationships between sign-based clusters and value-based clusters. Note that eigenvector components assuming different signs must assume different values, but the converse may not hold. We have the following:

Lemma 1: A sign-based cluster is either a value-based cluster or a collection of them where nodes share the same sign sequence in the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$.

Consider, for example, Figure 4. We note that sensors 0 through 5 form a sign-based cluster because they assume the same sign (not values) for each of the three eigenvectors. This sign-based cluster is a collection of three value-based clusters: sensors 0-1, 2-3 and 4-5.

Corollary 1: Any K eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, for $1 < K \leq rank(\mathbf{A}^{\mathrm{T}}\mathbf{A})$, can identify at least K sign-based clusters, which in turn can identify at least K collections of value-based clusters.

Proof: The result follows from Theorem 2 and Lemma 1. ■

By Corollary 1, we note that by reading off the component signs of the K eigenvectors, we can find at least K sign-based clusters and hence K collections of value-based clusters. However, the given K eigenvectors may specify more value-based clusters than sign-based clusters. For example, in the sensor-target scenario of Figure 3, the three eigenvectors of Figure 4 specify three value-based clusters for sensors 0-5 while exhibiting only one sign-based cluster consisting of these six sensors.

This means that although it is relatively easy to compute sign-based clusters, they may not reveal all the value-based clusters. We show below in Corollaries 2 and 3 some circumstances where this problem will not occur.

Corollary 2: Suppose that there are no more than K equivalence sets of nodes, for $1 < K \leq rank(\mathbf{A}^{\mathrm{T}}\mathbf{A})$. Then there are at most K value-based clusters, and sign-based clustering based on any K eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$, can identify all these value-based clusters.

Proof: The result follows from Theorem 1 and Corollary 1. ■

By Theorem 2, K eigenvectors can identify at least K sign-based clusters. In our experiments (see Section 6), it is often the case that K eigenvectors can identify more than K sign-based clusters. For instance, in the line detection example of Section 3, the five eigenvectors of Figure 7 identify six sign-based clusters. In this case, we note the following:

Corollary 3: Suppose that K eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ can identify S sign-based clusters, and there are E equivalence sets of nodes. Then sign-based clustering based on these K eigenvectors can identify $\min(S, E)$ value-based clusters or their collections.

Proof: The result follows from Theorem 1 and Corollary 1. ■

The line detection example of Section 3 illustrates Corollary 3. In this example, $K = 5$ eigenvectors in Figure 7 can identify $S = 6$ sign-based clusters and all the $E = 6$ value-based clusters in Figure 6.

Suppose that the number of equivalence sets is not larger than $rank(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ or the number of sign-based clusters identifiable by the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. Then, by Corollaries 2 and 3, we can compute all value-based clusters via sign-based clustering by just reading off the signs of components in the sign-based clusters.

For the case when the number of equivalence sets is larger than $rank(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ or the number of identifiable sign-based clusters, sign-based clusters can serve as initial approximations in search for value-based clusters via methods such as k-means. We discuss this in Section 5.

5. USE OF SIGN-BASED CLUSTERS AS INITIAL APPROXIMATIONS TO K-MEANS

By Lemma 1, a sign-based cluster specifies a value-based cluster or a collection thereof. Note also that k-means is supposed to cluster nodes which have similar component values in the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$. Hence it is natural to consider using sign-based clusters as initial approximations for k-means.

In fact, a pair of nodes in a sign-based cluster are expected to be closer to each other than a random pair of nodes. We can quantify this by considering the following two cases:

Case 1 (random node pair). Consider nodes a and b in a d-dimensional coordinate system which are drawn uniformly at random from inside a unit cube, that is, each coordinate of a or b is drawn uniformly random from the range $[-1, 1]$. Thus, the nodes lie in a hypercube of side length 2.

Case 2 (sign-based node pair). Consider any node pair a and b which have matching sign sequences in their coordinates. This means the two nodes are in the same quadrant of the d-dimensional coordinate system, and will be drawn randomly from a hypercube whose side has length 1.

Theorem 3: On average, the squared Euclidean distance between a and b is $(2/3)d$ and $d/6$ for Case 1 and 2, respectively.

Proof: Let a_i and b_i be the components of a and b, respectively. The results follow from expectation:

$$E\left[\sum_{i=1}^{d}(a_i - b_i)^2\right] = \sum_{i=1}^{d}(E[a_i^2] - 2E[a_i]E[b_i] + E[b_i^2])$$

■

Since $((2/3)d)/(d/6) = 4$, we can expect that a pair of nodes in a sign-based cluster is about 4 times closer in the squared Euclidean distance than a random pair. This means that a sign-based cluster is a good initial approximation in computing a value-based cluster or a collection thereof using heuristics such as k-means.

Note that when a finer grain sign-based clustering is used as an initial approximation, k-means faces less uncertainty in partitioning a relatively large collection of value-based clusters. For scenarios where there are sufficiently many clusters to be found, this can significantly shorten the running time of k-means.

Fortunately, in our experiments with the line detection exemplar problem, we have seen that the eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ can yield fine-grained sign-based clustering with a high probability. That is, K eigenvectors can identify substantially more than K sign-based clusters. Table 2 in Section 6 shows a summary of some of the findings.

We note that for the 4-line case, $K = 4$ eigenvectors can identify 8 sign-based clusters with a probability of 97%, and for the 5-line case, $K = 5$ eigenvectors can identify 14 or more clusters with a probability of 97.7%. When more sign-based clusters are identified, on the average each sign-based cluster becomes smaller, thereby achieving finer-grain sign-based clustering.

6. PERFORMANCE EVALUATION WITH SIMULATION

In this section we present a qualitative comparison of the performance of sign-based spectral clustering of this paper to a state-of-the-art k-means-based spectral clustering, based on numerical simulations. As we will see, sign-based clustering, in spite of being simpler and faster, can perform comparably to more complex techniques. In addition, sign-based clustering can provide good initial approximations to speed up the execution of the k-means algorithm.

6.1. Related work on k-means

A number of previous works have applied spectral techniques to deduce clustering in data of interest. In general, such works share the approach of first obtaining eigenvectors of some data relationship matrix, and then using some custom technique to compute clusters from the eigenvectors. For example, in an early work on image segmentation [6], researchers partitioned the input data repeatedly based on the values of the elements in the second eigenvector; the resulting recursive clustering was thus hierarchical. Subsequent work improved upon this by considering the top k eigenvectors of the data relationship matrix, instead of just the second. In particular, Ng et al. [7] applied the standard k-means clustering technique [11] to points in k-dimensional space whose coordinates are the respective elements of the top k eigenvectors. Applying k-means clustering to a low-dimensional eigenbasis instead of the input data directly allowed it to perform significantly better.

The heuristic algorithms used to implement k-means clustering have various limitations, such as a potentially superpolynomial running time [8], sensitivity to initial values, or the inability to deduce automatically the number of clusters. While researchers have tried to address some of these limitations in follow-on works, the solutions remain relatively complex [12].

The sign-based spectral clustering studied in this paper is a simple alternative to the k-means-based clustering algorithms; in particular, it does not depend on initial values, and does not require multiple iterations; in fact, after the eigenvectors are computed, its run-time of $O(Nd)$ is favorable to the $O(Ndk)$ run-time of a k-means iteration, where N is the number of data objects, d number of feature dimensions per object, and k number of clusters requested from the algorithm. Note that $O(Nd)$ or $O(Ndk)$ corresponds to the size of the data set that the sign-based spectral clustering or a k-means iteration needs to examine, respectively. The number of iterations k-means requires can be large and unpredictable, especially when the number of clusters is not known a priori, and as a result, an incorrect number of clusters is input to the algorithm.

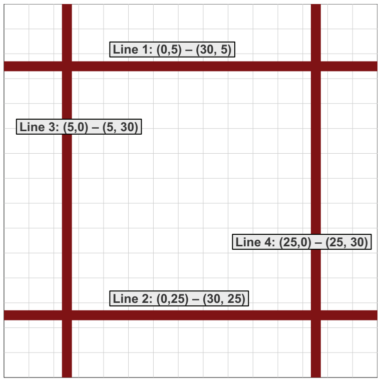


Fig. 8: A 4-line scenario used to compare sign-based and k-means clustering.

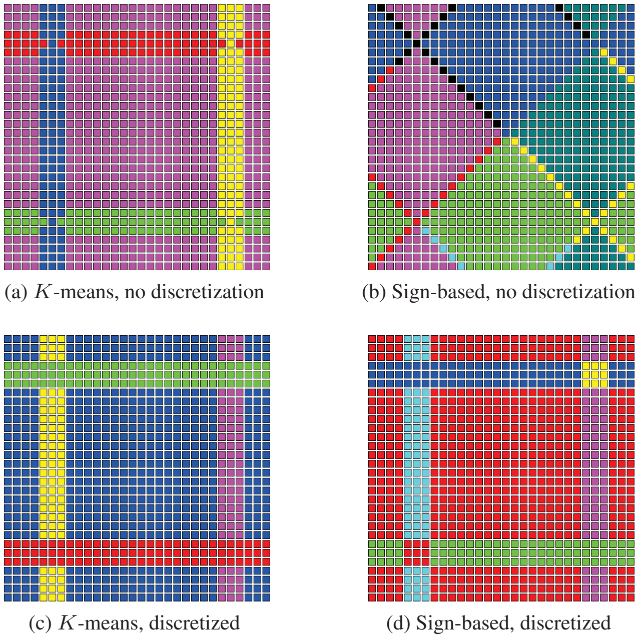
6.2. Performance Comparison in Robustness and Speed

We use a simple 4-line scenario, depicted in Figure 8, to compare k-means and sign-based clustering. We assume that there are 900 sensors arranged in a 30×30 grid surrounding the 4 line segments. We further model a sensor's measurement of a particular line by an inverse-square function of the sensor's distance from the nearest point on the line. Thus, based on these measurements, we form a 4×900 sensor-to-line input matrix $\mathbf{A}$. Lastly, for discrete cases we use a simple bi-level step function where sensors closer to the line than a threshold distance of 0.5 attain a measurement magnitude 1.0, and 0.01 otherwise.

We compare the performances of k-means-based and sign-based spectral clustering on raw and discrete inputs. The outcomes of the 4 possible scenarios appear in Figure 9. As we can see in Figure 9b, without discretization the sign-based clustering obtains a far worse solution than k-means, classifying mistakenly whole diamond-shaped regions of the background together with the points near target lines. However, with discretization, sign-based clustering finds an accurate solution. The solution, shown in Figure 9d, is comparable to k-means solutions. More precisely, we see that there are 6 sign-based clusters identified, corresponding to the 6 region colors in Figure 9d. Under the given discretization, a sensor's measurement of a line assumes only one of the two values 1 or 0.01, and there are only a total of 9 equivalence sets (four line stripes, four intersection blocks and one remaining background region). By Corollary 3, the $K = 4$ eigenvectors of $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ must identify min (6, 9) = 6 value-based clusters or their collections. The three left out are the ones corresponding to the smallest equivalence sets which are intersection blocks.

6.3. Evaluation of Using Sign-based Clusters as Initial Approximations for k-means Clustering

In Section 5 we argued that sign-based clusterings will likely be better initial approximations for the k-means algorithm than the randomly generated ones. In this section we report the results of numerical experiments in order to evaluate this proposition.



(a) K-means, no discretization

(b) Sign-based, no discretization

(c) K-means, discretized

(d) Sign-based, discretized

Fig. 9: A comparison of clustering outcomes under sign-based and k-means clustering, and with or without discretized input weights.

We used test cases similar to the test scenario of Section 6.2. We again assume 900 sensors arranged in a 30×30 grid. This time, we vary the number M of line segments being measured by sensors; furthermore, each line segment is determined by a pair of grid points chosen uniformly at random. Lastly, we modeled each sensor's measurement of a particular line by an inverse-square function of the sensor's distance from the nearest point on the line. From the measurements we form a $M \times 900$ sensor-to-line input matrix $\mathbf{A}$.

For each M-line test case, we performed two types of k-means clustering: that with randomly chosen initial centroid values, denoted "random" for short, and that which used the outcomes of sign-based clustering as its initial values, denoted "sign-based" for short. In both cases, we varied the desired number of clusters from M to 2^{M-1}; recall that the latter is the maximum number of clusters obtainable by sign-based clustering from $M - 1$ non-principal eigenvectors. In cases where sign-based clustering returned a larger than desired number of clusters, we merged an appropriate number of smallest clusters to reduce their count. We then computed a speedup as the ratio of the running times of sign-based and random k-means clustering, and obtained speedup statistics over 100 runs per test case. We show the results for $M = 4$ and 5 in Table 1. Note that the running time of k-means is sensitive to initial approximations. As a result, we see fluctuations in average speedup as #clusters increases in the 5 lines

Number of clusters	Avg. speedup (stdev)	Number of clusters	Avg. speedup (stdev)
3 lines		5 lines	
3	0.93 (0.28)	5	1.47 (0.52)
4	1.47 (0.49)	6	1.66 (0.53)
4 lines		7	1.59 (0.61)
4	1.19 (0.52)	8	1.35 (0.58)
5	1.59 (0.49)	9	1.27 (0.54)
6	1.53 (0.62)	10	1.34 (0.62)
7	1.32 (0.61)	11	1.10 (0.47)
8	1.32 (0.64)	12	1.17 (0.49)
		13	1.20 (0.44)
		14	1.17 (0.47)
		15	1.21 (0.44)
		16	1.41 (0.53)

Table 1: Measurements of speedup when using sign-based instead of random initial values for the k-means algorithm.

case. Nevertheless, we observe that using sign-based clusters as initial approximations lead to a noticeable overall speedup.

7. DISCUSSION AND CONCLUSION

Sign-based spectral clustering is simple, fast and robust. It is a direct method, requiring no heuristics in searching for clusters. Based on the orthogonality property of eigenvectors, we have shown that by using an increasing number of eigenvectors sign-based clustering can reveal clustering structures at increasingly refined granularity. The proof given in the paper appears to be new.

Sign-based clustering is especially useful when it can identify a relatively large numbers of clusters close to the target number of clusters. In this case, the resulting sign-based clusters may identify all or most of value-based clusters of interest, or they may serve as a good initial approximation for cluster finding heuristics such as k-means. This paper has provided some theoretical and empirical basis for our earlier experimental work in this area [13, 14]. However, results thus far should be regarded as preliminary; constrained by simulation time, at present our simulation results are limited to simple examples and a relatively small number of runs. Further investigation is needed before we can fully understand the potential and limitations of sign-based spectral clustering.

Number of clusters	Percentage	Avg. speedup (stdev)
3 lines		
4	100.0%	1.46 (0.50)
4 lines		
7	2.6%	1.40 (0.64)
8	97.4%	1.34 (0.58)
5 lines		
12	0.2%	1.08 (0.02)
13	2.0%	1.24 (0.64)
14	9.9%	1.21 (0.49)
15	35.5%	1.19 (0.47)
16	52.3%	1.20 (0.48)

Table 2: Breakdown of the number of clusters obtained through sign-based clustering of 3,4 or 5 line input cases.

8. REFERENCES

[1] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.

[2] S. White and P. Smyth, "A spectral clustering approach to finding communities in graphs," in *SIAM International Conference on Data Mining*, 2005.

[3] D. Higham, G. Kalna, and M. Kibble, "Spectral clustering and its use in bioinformatics," *Journal of Computational and Applied Mathematics*, vol. 204, no. 1, pp. 25–37, July 2007.

[4] M. Kurucz, A. Benczúr, K. Csalogány, and L. Lukács, "Spectral Clustering in Telephone Call Graphs," in *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop*, 2007.

[5] S. Foucher and L. Gagnon, "Automatic detection and clustering of actor faces based on spectral clustering techniques," in *CRV '07: Proceedings of the Fourth Canadian Conference on Computer and Robot Vision*, Washington, DC, USA, 2007, pp. 113–122, IEEE Computer Society.

[6] J. Shi and J. Malik, "Normalized cuts and image segmentation," in *IEEE Conf. Computer Vision and Pattern Recognition*, June 1997.

[7] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems (NIPS) 14*, 2002.

[8] D. Arthur and S. Vassilvitskii, "On the worst case complexity of the k-means method," in *22nd Annual ACM Symposium on Computational Geometry*, 2006.

[9] R. O. Duda and P. E. Hart, "Use of the Hough Transformation to Detect Lines and Curves in Pictures," *Communications of the ACM*, pp. 11–15, Jan. 1972.

[10] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, NY, 1979.

[11] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281–297.

[12] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *In Advances in Neural Information Processing Systems (NIPS) 17*, 2005.

[13] H. T. Kung and D. Vlah, "A Spectral Clustering Approach to Validating Sensors via Their Peers in Distributed Sensor Networks," in *Second International Workshop on Sensor Networks (SN2009)*, San Francisco, CA, Aug. 2009.

[14] H. T. Kung and D. Vlah, "Validating Sensors in the Field via Spectral Clustering Based on Their Measurement Data," in *Military Communications Conference (MILCOM 2009)*, Boston, MA, Oct. 2009.