Journal of Machine Learning Research 20 (2019) 1-56 Submitted 12/18; Revised 7/19; Published 10/19

Learning by Unsupervised Nonlinear Diffusion

Mauro Maggioni MAURO.MAGGIONI@JHU.EDU
Department of Mathematics, Department of Applied Mathematics and Statistics, Mathematical Institute of Data Sciences, Institute of Data Intensive Engineering and Science,
Johns Hopkins University, Baltimore, MD 21218, USA

James M. Murphy JM.MURPHY@TUFTS.EDU
Department of Mathematics
Tufts University, Medford, MA 02155, USA

Editor: Aapo Hyvärinen

Abstract

This paper proposes and analyzes a novel clustering algorithm, called *learning by unsupervised nonlinear diffusion (LUND)*, that combines graph-based diffusion geometry with techniques based on density and mode estimation. LUND is suitable for data generated from mixtures of distributions with densities that are both multimodal and supported near nonlinear sets. A crucial aspect of this algorithm is the use of time of a data-adapted diffusion process, and associated diffusion distances, as a scale parameter that is different from the local spatial scale parameter used in many clustering algorithms. We prove estimates for the behavior of diffusion distances with respect to this time parameter under a flexible nonparametric data model, identifying a range of times in which the mesoscopic equilibria of the underlying process are revealed, corresponding to a gap between within-cluster and between-cluster diffusion distances. These structures may be missed by the top eigenvectors of the graph Laplacian, commonly used in spectral clustering. This analysis is leveraged to prove sufficient conditions guaranteeing the accuracy of LUND. We implement LUND and confirm its theoretical properties on illustrative data sets, demonstrating its theoretical and empirical advantages over both spectral and density-based clustering.

Keywords: unsupervised learning, clustering, spectral graph theory, manifold learning, diffusion geometry

1. Introduction

Unsupervised learning is a central problem in machine learning, requiring that data be analyzed without a priori knowledge of any class labels. A common unsupervised problem is *clustering*, in which the data is to be partitioned into clusters so that each

cluster contains similar points and distinct clusters are sufficiently separated. Even with suitable definitions of "similarity" and "separation", this problem is typically ill-posed, requiring various geometric, analytic, topological, and statistical assumptions on the data and measurement method be imposed to make it tractable. Feature extraction is often combined with these standard methods (e.g. K-means) to improve clustering performance.

In particular, *spectral methods* construct graphs representing data, and use the spectral properties of the resulting graph Laplacian to produce structure-revealing features in the data. Graphs often encode pairwise similarities between points, typically at a local "spatial" scale, often determined by a parameter σ. For example only points x_i, x_j within distance 4σ of each other may be connected, with weight $\exp(-\|x_i - x_j\|_2^2/\sigma^2)$. From the graph, global features on the data may be derived, for example by considering the eigenfunctions of the random walk on the graph. Alternatively, graphs may be used to introduce data-adaptive distances, such as *diffusion distances*, which are associated to random walks and diffusion processes on graphs. Diffusion distances do not depend only on the graph itself, but also on a time parameter t that determines a scale *on the graph* at which these distances are considered, related to the time of diffusion or random walk. Choosing σ in graph-based algorithms, and both σ and t in the case of diffusion distances, is important in both theory and applications. However, their role is well-understood only in certain regimes (e.g. asymptotically for $\sigma, t \to 0^+$) which are of interest in some problems (e.g. manifold learning) but not necessarily for clustering.x

We propose the *Learning by Unsupervised Nonlinear Diffusion (LUND)* scheme for clustering, which combines diffusion distances and density estimation to efficiently cluster data generated from a nonparametric model. At the same time, we advance the understanding of the relationship between the local "spatial" scale parameter σ and the diffusion time parameter t in the context of clustering, demonstrating how the role of t can be exploited to successfully cluster data sets for which K-means, spectral clustering, or density-based clustering methods fail. We provide quantitative bounds and guarantees on the performance of the proposed clustering algorithm for data that may be highly nonlinear (i.e. non-convex, elongated, ellipsoidal, etc.) and of variable density.

1.1 Major Contributions and Outline

This article makes two major contributions. First, *explicit estimates on diffusion distances for nonparametric clustered data* are proved: we obtain lower bounds for the diffusion distance (see Definition 2.1, or Coifman et al. (2005)) between clusters, and upper bounds on the diffusion distance within clusters, as a function of the time parameter t and suitable properties of the clusters. These bounds yield a mesoscopic—not too small, not too large—diffusion time-scale at which diffusion distances separate clusters clearly and cohere points in the same cluster. These results, among other things, show how the role of the time parameter, which controls the scale "on the

data" of the diffusion distances, is very different from the commonly-used scaling parameter σ in the construction of the underlying graph, which is a local spatial scale measured in the ambient space. Relationships between t and σ are well-understood in the asymptotic case of $n \to +\infty$, $\sigma \to 0^+$ (at an appropriate rate with n; see Coifman et al. (2005), Lafon et al. (2006), and Von Luxburg (2007)) and $t \to 0^+$ (essentially Varadhan's lemma applied to diffusions on a manifold; see Den Hollander (2008), Jones et al. (2008), and references therein). These asymptotic relationships at small scales imply that the choice of t is essentially irrelevant, since in these limits diffusion distances are essentially geodesic distances. However, the clustering phenomena we are interested in are far from this regime, and we show that the interplay between t, σ, and n becomes crucial.

Second, *the LUND clustering scheme* is proposed and shown to enjoy performance guarantees for clustering of certain non-parametric mixture models. We prove sufficient conditions for LUND to correctly determine the number of clusters in the data and to have low clustering error. Computationally, we present an efficient algorithm implementing LUND, which scales near-linearly in the number of points n, in the ambient dimension D, and exponentially in the intrinsic dimension of the data. We verify the properties of the LUND scheme and algorithm on synthetic data, studying the relationships between the different parameters in LUND, in particular between σ and t, and compare with popular and related the graph-based *spectral clustering* and *fast search and find of density peaks clustering (FSFDPC)* (Rodriguez and Laio, 2014) algorithms, illustrating weaknesses of these methods and corresponding advantages of LUND. LUND may be understood as a combination of these two methods, in that it integrates diffusion distances (which are graph-based) and an outlier robustness procedure into the FSFDPC framework, which uses Euclidean distances. Indeed, our experiments illustrate how LUND combines the benefits of graph-based and density-based methods.

The outline of the article is as follows. Background is presented in Section 2. In Section 3, motivational data sets and a summary of the theoretical results are presented and discussed. Theoretical comparisons with related clustering methods are also made in Section 3. Estimates on diffusion distances are proved in Section 4. Performance guarantees for the LUND algorithm are proved in Section 5. Numerical experiments and computational complexity are discussed in Section 6. Conclusions and future research directions are given in Section 7.

2. Background

2.1 Background on Clustering

Given the wide variety of data of interest to scientific practitioners, many approaches to clustering have been developed, whose performance is often wildly variable and data-dependent.

2.1.1 K-Means

A classical and popular clustering algorithm is *K-means* (Steinhaus, 1957; Friedman et al., 2001) and its variants (Ostrovsky et al., 2006; Arthur and Vassilvitskii, 2007; Park and Jun, 2009), which is often used in conjunction with feature extraction methods. K-means partitions the data into K (a parameter) groups, $\{C_k\}_{k=1}^K$, chosen to minimize within-cluster dissimilarity: $C^* = \arg\min_{\{C_k\}_{k=1}^K} \sum_{k=1}^K \sum_{x \in C_k} \|x - \bar{x}_k\|_2^2$, where $\bar{x}_k$ is the mean of the k^{th} cluster (for a given partition, it is the minimizer of the least squares cost in the inner sum). While popular, K-means and its variants may perform poorly for data sets that are not the union of well-separated, near-spherical clusters, and are sensitive to outliers.

2.1.2 Hierarchical Clustering Methods

Hierarchical methods iteratively merge or split clusters in order to produce a multiscale family of partitions known as a dendrogram (Friedman et al., 2001). More precisely, a dendrogram for n data points is a family of clusterings $\{\mathcal{C}_i\}_{i=1}^n$ such that $\mathcal{C}_1$ is the clustering of n singleton clusters, and $\mathcal{C}_i$ is related to $\mathcal{C}_{i+1}$ in that the two clusters minimizing some linkage function in $\mathcal{C}_i$ are merged in $\mathcal{C}_{i+1}$. *Single linkage clustering (SLC)* (Sneath, 1957; Gower and Ross, 1969; Friedman et al., 2001) is a particular hierarchical clustering method that iteratively merges clusters according to the linkage function $\mathcal{L}_{\text{SLC}}(C_1, C_2) = \min_{x_1 \in C_1, x_2 \in C_2} \|x_1 - x_2\|_2$; metrics other than the ℓ^2 norm may be used. For clusters that are well-separated, single linkage clustering is known to perform well (Arias-Castro, 2011), despite lack of strong statistical consistency in dimensions greater than 1 (Hartigan, 1981).

2.1.3 Density and Mode-Based Methods

Density and mode-based clustering methods detect regions of high-density and low-density to determine clusters. The *DBSCAN* (Ester et al., 1996) and *DBCLASD* (Xu et al., 1998) algorithms assign to the same cluster points that are close and have many near neighbors, and flag as outliers points that lie alone in low-density regions. The *mean-shift* algorithm and variants (Fukunaga and Hostetler, 1975; Comaniciu and Meer, 2002; Chacón, 2012; Genovese et al., 2016) push points towards regions of high-density, and associate clusters with these high-density points, sometimes called *modes*. Both DBSCAN and mean-shift clustering suffer from a lack of robustness to outliers and depend strongly on parameter choices.

The *fast search and find of density peaks clustering algorithm* (FSFDPC) (Rodriguez and Laio, 2014) proposes to address these weaknesses. This method characterizes cluster modes as points that are far in Euclidean distance from points of higher density. FSFDPC has been widely applied (Spitzer et al., 2015; Wiwie et al., 2015; Sun et al., 2015; Rossant et al., 2016; Wang et al., 2016; Jia et al., 2016), and correctly clusters the data in Figure 1. However, we show that the standard FSFDPC method, while very popular in scientific applications, does not correctly cluster data unless

strong geometric and statistical assumptions on the data are satisfied. The main reason is that Euclidean distances are used to find modes, which is inappropriate for data drawn from mixtures of distributions supported near nonlinear sets (see, for example, Figure 18). Moreover, FSFDPC is not robust to outliers, which may be far from other points but be of low-density.

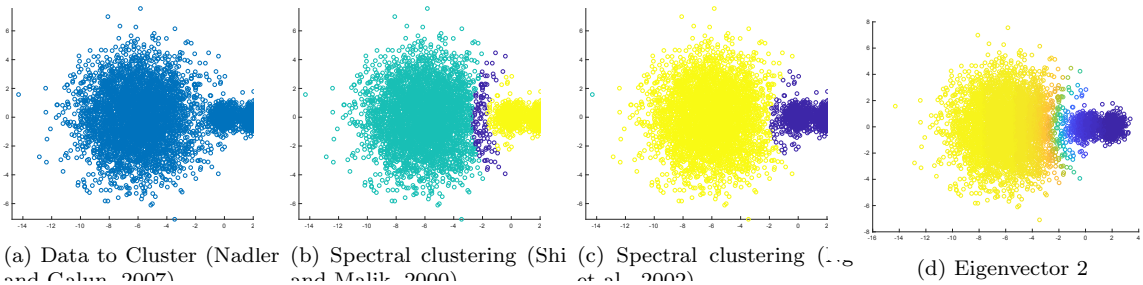
2.1.4 Spectral Methods

Spectral clustering methods compute features that reveal the structure of data that may deviate from the spherical, Gaussian shapes ideal for K-means, and in particular may be nonlinear or elongated in shape. This is done by building local connectivity graphs on the data that encode pairwise similarities between points, then computing a spectral decomposition of adjacency or random walk or Laplacian operators defined on this graph.

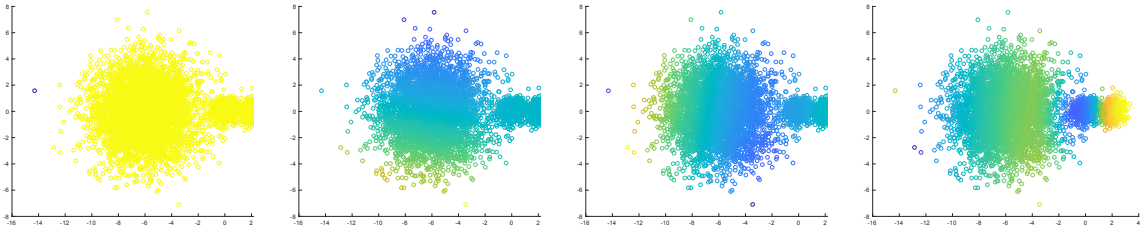
Let $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$ be a set of points to cluster. Let $\mathcal{G}$ be a graph with vertices corresponding to points of X and edges stored in an $n \times n$ symmetric weight matrix $\mathbf{W}$. Often one chooses $\mathbf{W}_{ij} = \mathcal{K}(x_i, x_j)$ for some (symmetric, often radial and rapidly decaying) nonnegative *kernel* $\mathcal{K} : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, such as $\mathcal{K}(x_i, x_j) = \exp(-||x_i - x_j||^2/\sigma^2)$ for some choice of scaling parameter $\sigma > 0$. The graph $\mathcal{G}$ may be fully connected, or it may be a nearest neighbors graph with respect to some metric. Let $\mathbf{D}$ be the diagonal matrix $\mathbf{D}_{ii} := \sum_{j=1}^n \mathbf{W}_{ij}$. The *graph Laplacian* is constructed as $\mathbf{L} = \mathbf{D} - \mathbf{W}$. One then normalizes $\mathbf{L}$ to acquire either the *random walk Laplacian* $\mathbf{L}_{\text{RW}} = \mathbf{D}^{-1}\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$ or the *symmetric normalized Laplacian* $\mathbf{L}_{\text{SYM}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}} = \mathbf{I} - \mathbf{D}^{-\frac{1}{2}}\mathbf{W}\mathbf{D}^{-\frac{1}{2}}$. We focus on $\mathbf{L}_{\text{SYM}}$ in what follows. It can be shown that $\mathbf{L}_{\text{SYM}}$ has real eigenvalues $0 = \lambda_1 \leq \cdots \leq \lambda_n \leq 2$ and corresponding eigenvectors $\{\phi_i\}_{i=1}^n$. The original data X can be clustered by clustering the embedded data $x_i \mapsto (\phi_1(x_i), \phi_2(x_i), \ldots, \phi_M(x_i))$ for an appropriate choice of $M \leq n$. In this step typically K-means is used, though Gaussian mixture models may (and perhaps should) be used, as they enjoy, unlike K-means, a suitably-defined statistical consistency guarantee in the infinite sample limit (Athreya et al., 2017).

Spectral clustering relaxes a graph-cut problem: for a collection of subsets $X_1, \ldots, X_K \subset X$, the corresponding *normalized cut* is $\text{Ncut}(X_1, \ldots, X_K) = \sum_{k=1}^K \text{cut}(X_k, X_k^c)/\text{vol}(X_k)$,where $\text{cut}(A, B) = \sum_{x_i \in A, x_j \in B} \mathbf{W}_{ij}$, $\text{vol}(A) = \sum_{x_i \in A} \sum_{j=1}^n \mathbf{W}_{ij}$. Minimizing Ncut yields clusters that are simultaneously separated and balanced (Shi and Malik, 2000). This NP-hard problem may be relaxed by analyzing the first K eigenvectors of $\mathbf{L}_{\text{SYM}}$ (Shi and Malik, 2000; Ng et al., 2002), or via a semidefinite programming problem (Ling and Strohmer, 2018).

Weaknesses of spectral clustering were scrutinized by Nadler and Galun (2007). They show the top eigenvectors of the random walk matrix $\mathbf{P}$—defined on $\{x_i\}_{i=1}^n$ sampled from $p(x)$ proportional to $e^{-U(x)/2}$ for some potential function $U(x)$—converge under a suitable scaling as $n \to \infty$ to the top eigenfunctions of the Fokker-Planck operator $\mathcal{L}\psi(x) = \Delta\psi - \nabla\psi \cdot \nabla U = -\mu\psi(x)$. The characteristic time scales of the stochastic differential equation (SDE) $\dot{x}(t) = -\nabla U(x) + \sqrt{2}\dot{w}(t), w$ a Wiener process, determine



(a) Data to Cluster (Nadler and Galun, 2007) (b) Spectral clustering (Shi and Malik, 2000) (c) Spectral clustering (Ng et al., 2002) (d) Eigenvector 2



(e) Eigenvector 3 (f) Eigenvector 4 (g) Eigenvector 5 (h) Eigenvector 6

Figure 1: In (a), three Gaussians of essentially the same density are shown. Results of spectral clustering are shown in (b) (Shi and Malik, 2000) and (c) (Ng et al., 2002). In (d) - (h), the first five non-trivial eigenvectors are shown. As noted by Nadler and Galun (2007), the underlying density for this data yields a Fokker-Planck operator whose low-energy eigenfunctions cannot distinguish between the two smaller clusters, thus preventing spectral clustering from succeeding: higher energy eigenfunctions are required. For this example, the sixth non-trivial eigenvector localizes sufficiently on the small clusters to allow for correct determination of the cluster structure; this eigenvector is not used in traditional spectral clustering algorithms.

the structure of the leading eigenfunctions of $\mathcal{L}$ (Gardiner, 2009): they correspond to the time scales of the slowest transitions between different clusters and the equilibrium times within clusters. The relationships between these quantities determine which eigenfunctions of $\mathcal{L}$ (or $\mathbf{P}$) reveal the cluster structure in the data. Gavish and Nadler (2013) further analyze related connections between normalized cuts and cluster exit times. Nadler and Galun (2007) present data which cannot be clustered with spectral clustering (Shi and Malik, 2000; Ng et al., 2002); see Figure 1.

2.2 Background on Diffusion Distances

One of the primary tools in the proposed clustering algorithm is *diffusion distances*, a class of data-dependent distances computed by constructing Markov processes on data that capture its intrinsic structure (Coifman et al., 2005; Coifman and Lafon, 2006; Lafon et al., 2006; Coifman et al., 2008; Singer and Coifman, 2008; Rohrdanz et al., 2011; Zheng et al., 2011; Lederman and Talmon, 2018; Lederman et al., 2015; Czaja et al., 2016; Li et al., 2017). We consider diffusion on the point cloud $X = \{x_i\}_{i=1}^n \subset \mathbb{R}^D$ via a Markov chain (Levin et al., 2009) with state space X. Let $\mathbf{P}$ be the corresponding $n \times n$ transition matrix. The following shall be referred to as the *usual assumptions* on $\mathbf{P}$: $\mathbf{P}$ is reversible, irreducible, aperiodic, and therefore ergodic. A common construction for $\mathbf{P}$, and the one we consider in the algorithmic sections of this article, is to first compute a *weight matrix* $\mathbf{W}$, where $\mathbf{W}_{ij} = \exp(-\|x_i - x_j\|_2^2/\sigma^2)$, $i \neq$

j for some appropriate scale parameter $\sigma \in (0, \infty)$. The parameter σ encodes the interaction radius of each point: σ large allows for long-range interactions between points that are ℓ^2-far, while σ small allows only for short-range interactions. Then $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{D}$ is the diagonal degree matrix with $\mathbf{D}_{ii} = \sum_{\ell=1}^{n} \mathbf{W}_{ij}$. This[1] row-normalizes $\mathbf{P}$: $\sum_{j=1}^{n} P_{ij} = 1, \ \forall i = 1, \ldots, n$. Since it is ergodic, $\mathbf{P}$ has a unique stationary distribution $\boldsymbol{\pi}$ satisfying $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$, given by $\boldsymbol{\pi}_i = \mathbf{D}_{ii} / \sum_{j=1}^{n} \mathbf{D}_{jj}$.
Diffusion processes on graphs lead to a data-dependent notion of distance, known as *diffusion distance* (Coifman et al., 2005; Coifman and Lafon, 2006). While the focus of the construction is on diffusion distances and the diffusion process itself, we mention that *diffusion maps* provide a way of efficiently computing and visualizing large-time diffusion distances in Euclidean space, and at the same time may be understood as a type of nonlinear dimension reduction, in which data in a high number of dimensions may be embedded in a low-dimensional space by a nonlinear coordinate transformation. In this regard, diffusion maps are related to nonlinear dimension reduction techniques such as Isomap (Tenenbaum et al., 2000), Laplacian eigenmaps (Belkin and Niyogi, 2003), and local linear embedding (Roweis and Saul, 2000), kernel PCA (with a data-adapted kernel), among many others.

Definition 2.1 *Let $X = \{x_i\}_{i=1}^{n} \subset \mathbb{R}^D$ and let $\mathbf{P}$ be a Markov process on X satisfying the usual assumptions and with stationary distribution $\boldsymbol{\pi}$. Let $\boldsymbol{\pi}_0$ be a probability distribution on X. For points $x_i, x_j \in X$, let $p_t(x_i, x_j) = (\mathbf{P}^t)_{ij}$, for some $t \in [0, \infty)$. The* diffusion distance at time t *between $x, y \in X$ is defined, for $\nu = \boldsymbol{\pi}_0 / \boldsymbol{\pi}$, by*

$$D_t(x, y) = \sqrt{\sum_{u \in X} \left(p_t(x, u) - p_t(y, u)\right)^2 \nu(u)} = ||p_t(x, \cdot) - p_t(y, \cdot)||_{\ell^2(\nu)} .$$

If the underlying graph is generated from data sampled from a low-dimensional manifold, then diffusion distance parametrizes this low-dimensional structure (Coifman et al., 2005; Jones et al., 2008; Singer et al., 2009; Singer and Wu, 2012, 2016; Talmon and Wu, 2018). Indeed, diffusion distances admit a formulation in terms of the (right) eigenfunctions of $\mathbf{P}$:

$$D_t(x, y) = \sqrt{\sum_{\ell=1}^{n} \lambda_\ell^{2t} (\psi_\ell(x) - \psi_\ell(y))^2}, \tag{2.2}$$

where $\{(\psi_\ell, \lambda_\ell)\}_{\ell=1}^{n}$ are the right eigenpairs of $\mathbf{P}$, ordered so that $1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_n > -1$, and noting that ψ_1 is constant by construction.
Diffusion distances are parametrized by t, which measures how long the diffusion process on $\mathcal{G}$ has run. Small t allows a small amount of diffusion, which may prevent the interesting geometry of X from being discovered, but provides detailed, fine scale

1. Note that with some abuse of notation we denote the entries of $\mathbf{P}$ by P_{ij}, reserving the notation $\mathbf{P}_{ij}$ for block submatrices of $\mathbf{P}$ that will be introduced and used later.

information. Large t allows the diffusion process to run for so long that the fine geometry may be washed out, leaving only coarse scale information. We will relate properties of clustered data X to t.

3. Data Model and Overview of Main Results

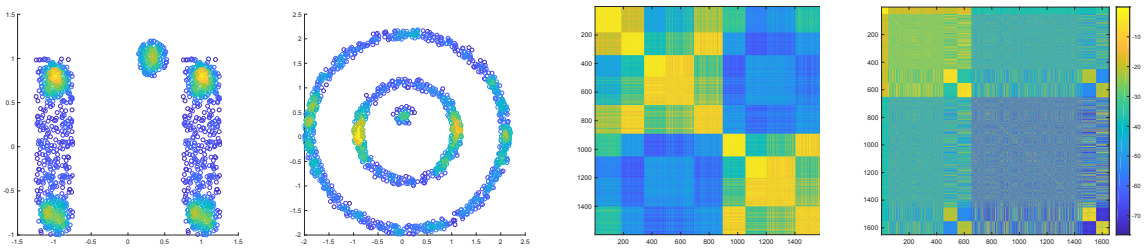
Among the main results of this article are sufficient conditions for clustering certain discrete data $X \subset \mathbb{R}^D$. The data X is modeled as a realization from a probability distribution

$$\mu = \sum_{k=1}^{K} w_k \mu_k, \ w_k \geq 0, \ \sum_{k=1}^{K} w_k = 1, \tag{3.1}$$

where each μ_k is a probability measure. Intuitively, our results require *separation and cohesion* conditions on $\{\mu_k\}_{k=1}^K$. That is, each μ_k is far from $\mu_{k'}$, $k \neq k'$ and connections are strong (in a suitable sense) within each μ_k. $X = \{x_i\}_{i=1}^n$ is generated by drawing, for each i, one of the K clusters, say k_i, according to the multinomial distribution with parameters $(w_1, \ldots, w_K)$, and then drawing x_i from μ_{k_i}. The clusters in the data are defined as the subsets of X whose samples were drawn from a particular μ_k, that is, we define the cluster $X_k := \{x_i \in X : k_i = k\}$. Given X, the goal of clustering is to estimate these X_k, and to estimate K. Throughout the theoretical analysis of this article, we will define the accuracy of a set of labels $\{Y_i\}_{i=1}^n$, $Y_i \in \{1, \ldots, K\}$, learned from an unsupervised algorithm to be $|\{i \mid Y_i = k_i\}|/n$, i.e. the proportion of points correctly labeled.

The model (3.1) is *nonparametric* and makes few explicit assumptions on μ. We will allow μ_k to be supported near a non-linear set (e.g. a nonconvex subset, or a submanifold in $\mathbb{R}^D$) and be multimodal (i.e. with multiple high-density regions). These features may cause prominent clustering methods to fail, e.g. K-means, which requires near-spherical or well-separated clusters (Dasgupta and Schulman, 2007; Mixon et al., 2017); Gaussian mixture models, which handle well spherical and ellipsoidal clusters, but may struggle with clusters exhibiting different, non-elliptical geometries (for example those shown in Figure 2 (b)); spectral clustering, which can fail for highly elongated clusters or clusters of different sizes and densities; density-based methods, which are sensitive to noise and clusters of very different densities, and may require care in setting parameters or implementing adaptive parameters. Two simple, motivating examples are in Figure 2. They feature variable densities, variable levels of connectivity, both within and across clusters, and (for the second example) nonlinear cluster shapes.

The estimates for the behavior of diffusion distances that we derive will then be leveraged to prove that the LUND clustering scheme correctly labels the points and also correctly estimates the number of clusters, while other clustering schemes fail to cluster these data sets correctly.



(a) Bottleneck data (b) Nonlinear data (c) $\mathbf{P}$ for data in (a). (d) $\mathbf{P}$ for data in (b).

Figure 2: Subfigures (a) and (b) show two data sets—linear and nonlinear—colored by $\log_{10}(p(x))$, where $p(x)$ is the empirical density. In (c), (d), we show the corresponding Markov transition matrices $\mathbf{P}$, with entry magnitudes shown in $\log_{10}$ scale. We sorted the rows and columns so that the structure in these matrices become apparent (of course, the algorithms are independent of the sorting). The Markov chains are ergodic, but close to being reducible. Indeed, these transition matrices show hierarchical structure, with large within-cluster transitions and uniformly small probabilities of transition between clusters. The analysis of Section 4 makes this intuition precise. The transition matrices were constructed using the Gaussian kernel as in Section 2.2, with Euclidean distances and $\sigma = .25$

3.1 Summary of Main Results

Our first result shows that the within-cluster and between-cluster diffusion distances can be controlled, as soon as $\mathbf{P}^t$ is approximately block constant in some sense. Define the (worst-case) within-cluster and between-cluster diffusion distances as:

$$D_t^{\text{in}} = \max_k \max_{x,y \in X_k} D_t(x,y), \quad D_t^{\text{btw}} = \min_{k \neq k'} \min_{x \in X_k, y \in X_{k'}} D_t(x,y)\,. \tag{3.2}$$

Our results guarantee that D_t^{in} is small and D_t^{btw} is large in terms of a constant ϵ, for all t in some time interval $\mathcal{T}$. The interval $\mathcal{T}$ depends on both data-driven quantities of the matrix $\mathbf{P}$ (which may be understood as geometrically intrinsic to the underlying cluster structure), as well as ϵ. This simplification of Theorem 4.12 holds:

Theorem 3.3 *Let $X = \bigcup_{k=1}^K X_k$ and let $\mathbf{P}$ be a corresponding Markov transition matrix on X, inducing diffusion distances D_t. Then there exist non-negative constants $\{C_i\}_{i=1}^5$ such that the following holds: for any $\epsilon > 0$ we have*

$$C_1 \ln\left(\frac{C_2}{\epsilon}\right) < t < C_3\epsilon \quad \Longrightarrow \quad D_t^{in} \leq C_4\epsilon, \quad D_t^{btw} \geq C_5 - C_4\epsilon.$$

The constants $\{C_i\}_{i=1}^5$ are defined precisely in Section 4, but to get a sense of them, let $\mathbf{S}$ be an idealized version of $\mathbf{P}$ in which all edges between clusters are deleted and redirected back into the cluster; this will be made precise by the notion of stochastic complement in Definition 4.7. Let $\mathbf{S}^\infty$ be a block diagonal matrix consisting of K rank 1 blocks, corresponding to the equilibrium transition matrices of the the blocks of $\mathbf{S}$. The constants $\{C_i\}_{i=1}^5$ may be interpreted as follows:

- C_1, C_2 are related to the compactness or irreducibility of the clusters X_k, in terms of the rank of each block of $\mathbf{S}$. In particular, if $\mathbf{S}$ is constant on each

diagonal block corresponding to a cluster, then $C_1 = 0$ and $C_2 = O(1)$, independently of n.

- C_3 is related to the transition probabilities between clusters in $\mathbf{P}$. In the idealized case that there are no transitions between any of the clusters, $C_3 = \infty$. More generally, if the probabilities of transitions between clusters are small, then C_3 will be large.

- C_4 is related to the balance of $\mathbf{S}^\infty$ relative to $\mathbf{P}$. If each row of $\mathbf{S}^\infty - \mathbf{P}$ is constant, then $C_4 = 1/\sqrt{n}$.

- C_5 is related to how balanced the equilibrium distributions are on each cluster. If $\mathbf{S}^\infty$ is block constant with blocks of the same size, then $C_5 = 1/\sqrt{n}$.

The time interval $\mathcal{T} := [C_1 \ln\left(\frac{C_2}{\epsilon}\right), C_3\epsilon]$ depends on ϵ, and there is tension in the role of ϵ between the condition $t \in \mathcal{T}$ and the conclusion that $D_t^{\text{in}} \leq C_4\epsilon, \quad D_t^{\text{btw}} \geq C_5 - C_4\epsilon$. For large ϵ, $\mathcal{T}$ may be quite wide, but the conclusion on diffusion distances is weak (or even trivial if $\epsilon > C_5/C_4$). On the other hand, ϵ small induces strong separation between within-cluster diffusion distances and between-cluster diffusion distances, at the expense of shrinking $\mathcal{T}$. Indeed, for $0 < C_1, C_2, C_3 < \infty$ fixed, $\mathcal{T}$ shrinks to the empty set as $\epsilon \to 0^+$.
Suppose indeed that the each cluster is compact (C_1 small, $C_2 = O(1)$), the clusters are well-separated (C_3 large), and balanced in the sense that $\mathbf{S}^\infty - \mathbf{P}$ is constant and $\mathbf{S}^\infty$ is block-constant with blocks of the same size ($C_4 = C_5 = 1/\sqrt{n}$). Then for $\epsilon > 0$,

$$t \in \left[C_1 \ln\left(\frac{C_2}{\epsilon}\right), C_3\epsilon\right] \implies \frac{D_t^{\text{btw}}}{D_t^{\text{in}}} \geq \frac{C_5 - C_4\epsilon}{C_4\epsilon} = O\left(\frac{1}{\epsilon}\right),$$

which suggests strong separation with diffusion distances independently of n when ϵ is small. In particular, as the clusters become more separated (C_3 increases) the time interval on which the $D_t^{\text{btw}}/D_t^{\text{in}}$ remains large widens on the right. Similarly, the more compact or irreducible the clusters become (C_1 becomes smaller), the wider the interval becomes on the left. In the ideal case that $C_1 = 0$ (clusters are localized at single points), $C_3 = \infty$ (infinite separation between clusters), the ratio $D_t^{\text{btw}}/D_t^{\text{in}} = \infty$ for all t, due to the fact that D_t^{btw} is bounded away from 0 for all t while $D_t^{\text{in}} = 0$. We remark that as long as ϵ can be taken sufficiently small, due to the geometric properties of the data as determined by C_1, C_2, C_3, the lower bound on D_t^{btw} is positive. In the idealized case when $C_4 = C_5 = 1/\sqrt{n}$, then the lower bound is positive as long as $\epsilon < 1$. We note that C_1, C_2, C_3 are close to geometrically intrinsic, as will be discussed in Section 4.1.
The *LUND scheme* characterizes modes of the clusters $\{X_k\}_{k=1}^K$ as high-density points that are far in diffusion distance from other points of high-density, regardless of the shape of the support of the distribution. Once modes are learned, remaining points are subsequently assigned to a mode in an iterative fashion. To detect the modes,

we introduce two quantities: $p(x)$ is related to data density, while $\rho_t(x)$ is related to diffusion geometry.
Let p be a kernel density estimator (KDE) on X, for example $p(x) = \frac{1}{Z}\sum_{y\in NN(x)} \exp(-\|x-y\|_2^2/\sigma_0^2)$, for some choice of σ_0 and set of nearest neighbors $NN(x)$, normalized by Z so that $\sum_{x\in X} p(x) = 1$. Given D_t defined on X, let

$$\rho_t(x) = \begin{cases} \min\limits_{y\in X}\{D_t(x,y) \mid p(y) \geq p(x)\}, & x \neq \arg\max\limits_{y\in X} p(y), \\ \max\limits_{y\in X} D_t(x,y), & x = \arg\max\limits_{y\in X} p(y). \end{cases} \tag{3.4}$$

The function ρ_t measures the diffusion distance of a point to its D_t-nearest neighbor of higher empirical density. LUND proceeds by analyzing

$$\mathcal{D}_t(x) = p(x)\rho_t(x),$$

which is large only for high-density points that are far from their nearest diffusion neighbor of higher density. The function $\mathcal{D}_t$ serves two important purposes. First, its decay estimates the number of clusters in the data. Indeed, as will be shown in Section 5, under a flexible data model, $\mathcal{D}_t$ has K very large values with the rest very small, where K is the number of latent clusters. Second, the modes of the data are estimated as the maximizers of $\mathcal{D}_t$. Once these modes have been learned, they are given unique labels. Then, in order of decreasing density, points are assigned the same label as their D_t-nearest neighbor of higher density. In this sense, the labels of the modes are distributed—from high to low density—to the rest of the data. The LUND algorithm is detailed in Algorithm 1. A simpler variant of LUND, when K is known a priori, is detailed in Algorithm 2.
Functions of $p(x), \rho_t(x)$ other than multiplication could be used to construct $\mathcal{D}_t(x)$. The primary reason to consider the multiplication of these factors is to gain robustness to outliers. Indeed, an outlying point x_o may be very far from all other points in diffusion distance, simply because its Euclidean coordinates are very far from the rest of X. In this case, one would have $\rho_t(x_o)$ very large, and $p(x_o)$ very small. By constructing $\mathcal{D}_t$ as the product of the density and diffusion geometric measurements, $\mathcal{D}_t(x_o) = p(x_o)\rho_t(x_o)$ is not large (under a suitable regime of variation between $p(x_o)$ and $\rho_t(x_o)$), ensuring that a far outlier is not be selected as a data mode. More precisely, suppose that diffusion distances and $p(x)$ are computed with the same choice of scaling parameter σ and collection of nearest neighbors, so that the stationary distribution of $\mathbf{P}$ is equal to p: for all $x \in X$, $p(x) = \boldsymbol{\pi}(x)$. Suppose that X is fixed except for the outlier point x_o, which we assume to be the element of X with lowest empirical density. Letting x_o^{NN} be the nearest neighbor of x_o of higher empirical density and $\boldsymbol{\pi}_0$ an arbitrary initial distribution,

$$\mathcal{D}_t(x_o) = \rho_t(x_o)p(x_o) = \sqrt{\sum_{\ell=1}^{n} \left(p_t(x_o,x_\ell) - p_t(x_o^{NN},x_\ell)\right)^2 \frac{\boldsymbol{\pi}_0(x_\ell)\boldsymbol{\pi}(x_o)^2}{\boldsymbol{\pi}(x_\ell)}} \leq \sqrt{n\boldsymbol{\pi}(x_o)}.$$

Algorithm 1 Learning by Unsupervised Nonlinear Diffusion (LUND) Algorithm

Input: X (data), σ_0 (kernel density bandwidth), σ (diffusion scaling parameter), t (time parameter), τ (threshold)

Output: Y (cluster assignments), $\hat{K}$ (estimated number of clusters)

Build Markov transition matrix $\mathbf{P}$ using scale parameter σ.
Compute KDE $p(x)$ for all $x \in X$ using kernel bandwidth σ_0.
Compute $\rho_t(x)$ for all $x \in X$.
Compute $\mathcal{D}_t(x) = \rho_t(x)p(x)$ for all $x \in X$.
Sort X according to $\mathcal{D}_t(x)$ in descending order as $\{x_{m_i}\}_{i=1}^n, n = |X|$.
Compute $\hat{K} = \inf\left\{k \;\middle|\; \frac{\mathcal{D}_t(x_{m_k})}{\mathcal{D}_t(x_{m_{k+1}})} > \tau\right\}$.
Assign $Y(x_{m_i}) = i,\ i = 1, \ldots, \hat{K}$, and $Y(x_{m_i}) = 0,\ i = \hat{K}+1, \ldots, n$.
Sort X according to $p(x)$ in decreasing order as $\{x_{\ell_i}\}_{i=1}^n$.
for $i = 1 : n$ **do**
  if $Y(x_{\ell_i})$=0 **then**
    $Y(x_{\ell_i}) = Y(x^*),\ x^* = \arg\min_y\{D_t(x_{\ell_i}, y) \mid p(y) \geq p(x_{\ell_i})$ and y is labeled$\}$.
  end if
end for

Noting $\lim_{\|x_o\|_2 \to \infty} \boldsymbol{\pi}(x_o) = 0$, it follows that $\lim_{\|x_o\|_2 \to \infty} \mathcal{D}_t(x_o) = 0$, so that as outliers move farther away from the rest of the data, they become less likely to be detected as modes. We remark that if outlier detection is performed on the data as a pre-processing step, this problem is less significant, since densities become more comparable across points. In this case, other constructions for $\mathcal{D}_t$ may be sufficiently robust, for example constructions that are additive in p, ρ_t.

The LUND algorithm combines density estimation (as captured by p) with diffusion geometry (as captured by ρ_t). The crucial parameter of LUND is the time parameter, which determines the diffusion distance D_t used. Theorem 3.3 may be used to show that there is a range of t for which applying the proposed LUND algorithm is provably accurate. The first concern is to understand conditions guaranteeing these modes are estimated accurately, the second that all other points are consequently labeled correctly. The following result summarizes Corollaries 5.4, 5.5, corresponding to the case when K is unknown a priori and must be estimated (as in Algorithm 1) or is known a priori (as in Algorithm 2).

Theorem 3.5 *Suppose $X = \bigcup_{k=1}^K X_k$ as above. Let $\mathcal{M}$ be the set of cluster density maxima:*

$$\mathcal{M} = \{p(x) \mid \exists k \in \{1, 2, \ldots, K\} \text{ such that } x = \arg\max_{y \in X_k} p(y)\}.$$

(a) Let $\{x_{m_i}\}_{i=1}^n$ be the points $\{x_i\}_{i=1}^n$, sorted so that $\mathcal{D}_t(x_{m_1}) \geq \mathcal{D}_t(x_{m_2}) \geq \cdots \geq \mathcal{D}_t(x_{m_n})$. Then Algorithm 1 correctly estimates K and labels all points correctly

Algorithm 2 LUND Algorithm, K Known

Input: X (data), σ_0 (kernel density bandwidth), σ (diffusion scaling parameter), t (time parameter), K (number of clusters)
Output: Y (cluster assignments)

Build Markov transition matrix $\mathbf{P}$ using scale parameter σ.
Compute a KDE $p(x)$ for all $x \in X$ using kernel bandwidth σ_0.
Compute $\rho_t(x)$ for all $x \in X$.
Compute $\mathcal{D}_t(x) = \rho_t(x)p(x)$ for all $x \in X$.
Sort X according to $\mathcal{D}_t(x)$ in descending order as $\{x_{m_i}\}_{i=1}^n, n = |X|$.
Assign $Y(x_{m_i}) = i,\ i = 1, \ldots, K$, and $Y(x_{m_i}) = 0,\ i = K+1, \ldots, n$.
Sort X according to $p(x)$ in decreasing order as $\{x_{\ell_i}\}_{i=1}^n$.
for $i = 1 : n$ **do**
  if $Y(x_{\ell_i})$=0 **then**
    $Y(x_{\ell_i}) = Y(x^*),\ x^* = \arg\min_y\{D_t(x_{\ell_i}, y) \mid p(y) \geq p(x_{\ell_i})$ and y is labeled$\}$.
  end if
end for

for any τ satisfying

$$\frac{\max(\mathcal{M})}{\min(\mathcal{M})}\frac{\max_{i=1,\cdots,K}\rho_t(x_{m_i})}{\min_{i=1,\cdots,K}\rho_t(x_{m_i})} < \tau < \frac{\min(\mathcal{M})}{\max(\mathcal{M})}\frac{D_t^{btw}}{D_t^{in}}. \tag{3.6}$$

(b) If K is known a priori, then Algorithm 2 labels all points accurately provided that

$$\frac{D_t^{in}}{D_t^{btw}} < \frac{\min(\mathcal{M})}{\max(\mathcal{M})}. \tag{3.7}$$

Theorem 3.3 suggests that the conditions (3.6), (3.7) will hold for a wide range of t, depending on $\max(\mathcal{M})/\min(\mathcal{M})$ and the underlying data geometry. Indeed, in the case that $C_4 = C_5 = 1/\sqrt{n}$, and the clusters are irreducible (C_1 small) and well-separated (C_3 large), the time interval guaranteeing $D_t^{\text{in}}/D_t^{\text{btw}} < \epsilon$ is $[C_1 \ln(\frac{C_2}{\epsilon}), C_3\epsilon]$, which is wide even for ϵ small. Indeed, setting $\epsilon = \min(\mathcal{M})/(2\max(\mathcal{M}))$ is sufficient to guarantee (3.7) holds for a wide range of t (always for C_1 sufficiently small and C_3 sufficiently large). Of course, the smaller the ratio $\min(\mathcal{M})/\max(\mathcal{M})$, the harder (3.7) is to satisfy.
Together with Theorem 3.5, this implies the proposed method (Algorithm 1) correctly labels the data and estimates the number of clusters K correctly. Note that (3.7) implicitly relates the density of the separate clusters to their geometric properties. Indeed, if the clusters are well-separated and cohesive enough, then $D_t^{\text{in}}/D_t^{\text{btw}}$ is very small, and a large discrepancy in the density of the clusters can be tolerated in inequality (3.7). Note that $D_t^{\text{in}}, D_t^{\text{btw}}, \min(\mathcal{M})$, and $\max(\mathcal{M})$ are invariant to increasing n, as long as the scale parameter in the kernel used for constructing diffusion distances and the KDE adjusts according to standard convergence results for graph

Laplacians (Belkin and Niyogi, 2005, 2007; Garcia Trillos et al., 2016, 2018). In this sense these quantities are properties of the mixture model (3.1), and neither of n nor the scale parameter of the diffusion process σ.
We note moreover that Theorem 3.5 suggests t must be taken in a mesoscopic range, that is, sufficiently far from 0 but also bounded. Indeed, for t small, D_t^{in} is not necessarily small, as the Markov process has not mixed locally yet. For t large, $\mathbf{P}^t$ is close to global stationarity, and the ratio $D_t^{\text{in}}/D_t^{\text{btw}}$ is not necessarily small, since D_t^{btw} will be small. In this case, clusters would only be detectable based on density, requiring thresholding, which is susceptible to spurious identification of regions around local density maxima as clusters.
We remark that a LUND prototype adapted to image data was proposed for the empirical study of high-dimensional hyperspectral images by Murphy and Maggioni (2018a,b, 2019b), where it is shown to enjoy competitive performance with state-of-the-art clustering algorithms on specific data sets. The LUND algorithm presented in the present work is more general and, in contrast with earlier related methods, appropriate for general point cloud data, not just images.

3.2 Comparisons with Related Clustering Algorithms

LUND combines graph-based methods with density-based methods, and it is therefore natural to compare it with spectral clustering and FSFDPC among other methods.

3.2.1 Comparison with Spectral Clustering

The normalized graph-cut problem in spectral clustering is related to the probability of transitioning between clusters in *one* time step (Meila and Shi, 2001). LUND uses intermediate time scales to separate clusters, namely the time scale at which the random walk has almost reached the stationary distribution conditioned on not leaving a cluster, and has not yet transitioned (with sizable probability) to a different cluster.
Spectral clustering enjoys performance guarantees under a range of model assumptions (Chen and Lerman, 2009a,b; Arias-Castro, 2011; Arias-Castro et al., 2011; Vidal, 2011; Zhang et al., 2012; Elhamifar and Vidal, 2013; Wang et al., 2015; Soltanolkotabi et al., 2014; Arias-Castro et al., 2017; Little et al., 2017). Under nonparametric assumptions on (3.1) with $K = 2$, Shi et al. (2009) show that the principal eigenfunctions and eigenvalues of the associated kernel operator $\mathcal{K}(f)(x) = \int K(x,y)f(y)d\mu(y)$ are closely approximated by the principal spectra of the kernel operators $\mathcal{K}_i(f)(x) = \int K(x,y)f(y)d\mu_i(y), i = 1,2$, possibly mixed up, depending on the spectra of $\mathcal{K}_1, \mathcal{K}_2$ and the weights w_1, w_2. This allows for the number of classes to be estimated accurately in some situations, and for points to be labeled by determining which distribution certain eigenvectors come from.
The related work of Schiebinger et al. (2015) provides sufficient conditions under the nonparametric model (3.1) for the low-dimensional embedding of spectral clus-

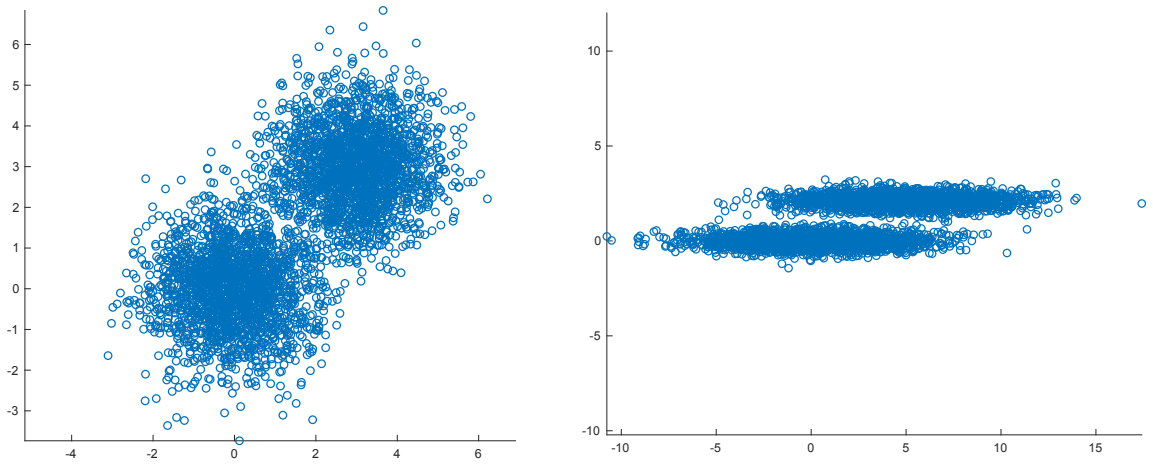
tering to map well-separated, coherent regions in input space to approximately orthogonal regions in the embedding space. This in turn implies that K-means clustering succeeds with high probability, thereby yielding guarantees on the accuracy of spectral clustering. These results depend on two quantities: with μ as in (3.1) and $\mathcal{K}$ a kernel, they define separation and cohesion quantities, respectively, as $\mathcal{S}(\mu) = \max_{i \neq j} \mathcal{S}(\mu_i, \mu_j), \Gamma_{\min}(\mu) = \min_{i=1,\ldots,K} \Gamma(\mu_i)$, where

$$\mathcal{S}(\mu_i, \mu_j) = \frac{1}{p(X)} \int_X \int_X \mathcal{K}(x,y) d\mu_i(x) d\mu_j(y), \Gamma(\mu_i) = \inf_{S \subset X} \frac{p(X)}{p(S)p(S^c)} \int_S \int_{S^c} \mathcal{K}(x,y) d\mu_i(x) d\mu_i(y),$$

$p(S) = \int_S \int_X \mathcal{K}(x,y) d\mu(x) d\mu(y)$. A major result of Schiebinger et al. (2015) is that spectral clustering is accurate with high probability depending on a confidence parameter β and the number of data samples n if

$$\frac{\sqrt{K(\mathcal{S}(\mu) + \mathcal{C}(\mu))}}{\min_{i=1,\ldots K} w_i} + \left(\frac{1}{\sqrt{n}} + \beta \right) \lesssim \Gamma^4_{\min}(\mu)\,, \tag{3.8}$$

where $\mathcal{C}(\mu)$ is a "coupling parameter" that is not germane to the present discussion. Condition (3.8) holds when the within-cluster coherence $\Gamma_{\min}(\mu)$ is large relative to the similarity between clusters $\mathcal{S}(\mu)$. Fixing the separation $\Gamma_{\min}(\mu)$, (3.8) is more likely to hold if the clusters are relatively spherical in shape. For example, in Figure 3 we represent two data sets, each consisting of two clusters, with comparable $\mathcal{S}(\mu)$, but substantially different $\Gamma_{\min}(\mu)$. Also note that in the finite sample case when $\frac{1}{\sqrt{n}}$ in (3.8) is non-negligible, the importance of $\Gamma_{\min}$ being not too small increases. The



(a) $\mathcal{S}(\mu) = 0.0533, \Gamma_{\min}(\mu) \approx 0.9550$.

(b) $\mathcal{S}(\mu) = 0.0523,\ \Gamma_{\min}(\mu) \approx 0.2560$.

Figure 3: In (a) and (b) two different mixtures of Gaussians are shown. The two mixtures have roughly the same measure of between-cluster distance $\mathcal{S}(\mu)$, but significantly different within-cluster coherence $\Gamma_{\min}(\mu)$. Spectral clustering will enjoy much stronger performance guarantees, according to Schiebinger et al. (2015), for the data in (a) compared to the data in (b), for a range of relevant choices of the parameter σ.

It is of related interest to compare LUND to spectral clustering by recalling (2.2). In the generic case that $\lambda_2 > \lambda_3$, the $(\psi_2(x) - \psi_2(y))^2$ term dominates asymptotically as $t \to \infty$. Hence, as $t \to \infty$, LUND bears resemblance to spectral clustering with the second eigenvector alone (Shi and Malik, 2000). On the other extreme, for $t = 0$, diffusion distances depend on all eigenvectors equally. Using the first K or the 2^{nd} through $(K+1)^{st}$ eigenvectors ψ_l is the basis for many spectral clustering algorithms (Ng et al., 2002; Schiebinger et al., 2015), and is comparable to LUND for $t = 0$, combined with a truncation of (2.2). Note that clustering with the kernel $\mathcal{K}$ alone relates to using all eigenvectors and $t = 1$. By allowing t to be a tunable parameter, LUND interpolates between the extremes of the K principal eigenvectors equally ($t = 0$ and truncating the eigendecomposition after the K^{th} or $(K+1)^{st}$ eigenvector), using the kernel matrix ($t = 1$), and using only the second eigenvector ($t \to \infty$). The results of Section 6 validate the importance of this flexibility.

An additional challenge when using spectral clustering is to robustly estimate K. The *eigengap* $\hat{K} = \arg\max_i \lambda_{i+1} - \lambda_i$ is a commonly used heuristic, but is often ineffective when Euclidean distances are used in the case of non-spherical clusters (Arias-Castro, 2011; Little et al., 2017). In contrast, Theorem 3.5 suggests LUND can robustly estimate K, which is shown empirically for synthetic data in Section 6.

It is also of interest to compare the guarantees of Theorem 3.3 to the analysis of spectral clustering of Nadler and Galun (2007); see Section 2.1.4. In particular, the guarantees of LUND require balancing two quantities: the within cluster mixing times and the between-cluster transition probabilities. These are analyzed precisely in Section 4, and are quantified in Theorem 3.3 by C_1 (within cluster mixing propensity) and C_3 (between cluster transition propensity). In the framework of continuous Fokker-Planck equations, these notions are closely related to relaxation time and first exit time, respectively. Nadler and Galun (2007) argue that as long as the first passage exit time is greater than the relaxation time within a cluster, for all clusters, then spectral clustering has a hope of achieving good results. The LUND algorithm relies on a similar fundamental observation, and the delicate balance between these two notions (within cluster mixing and between cluster transitions) are analyzed in a precise, quantitative sense for discrete Markov chains in Section 4, leading in Theorem 4.12 to a guarantee on the behavior of diffusion distances in terms of these quantities.

Computationally, LUND and spectral methods are essentially the same, with the bottleneck in complexity being either the spectral decomposition of a dense $n \times n$ matrix ($O(Mn^2)$ where M is the number of eigenvectors sought), or the computation of nearest neighbors when using a sparse diffusion operator or Laplacian (using an indexing structure for a fast nearest neighbors search, this is $O(C^d Dn \log(n))$, where d is the intrinsic dimension of the data).

3.2.2 Comparison With Local Graph Cutting Algorithms

The LUND algorithm bears some resemblance to local graph cutting algorithms (Spielman and Teng, 2004; Andersen et al., 2006, 2008; Andersen and Peres, 2009;

Spielman and Teng, 2013, 2014; Yin et al., 2017; Fountoulakis et al., 2017). These methods compute a cluster C around a given vertex v such that the conductance of C is high (see Definition 4.3), and which can be computed in sublinear time with respect to the total number of vertices in the graph n, and in linear time with respect to $|C|$. In order to avoid an algorithm that scales linearly (or worse) in n, global features—such as eigenvectors of a Markov transition matrix or graph Laplacian defined on the data—must be avoided. The Nibble algorithm (Spielman and Teng, 2013) and related methods (Andersen and Peres, 2009) compute approximate random walks for points nearby v, and truncate steps that take the random walker too far from already explored points. This accounts for the most important steps a random walker would take, and avoids considering all n vertices of the graph. In this sense, Nibble and related methods focus on local diffusion in order to compute a local cluster around a prioritized vertex v, while LUND focuses on both finding good starting points and a globally consistent partition of the whole graph, using nonlinear, typically large-time diffusion to uncover multitemporal structure.

3.2.3 Comparison with FSFDPC

The FSFDPC algorithm (Rodriguez and Laio, 2014) learns the modes of clusters in a manner similar to the method proposed in this article. In FSFDPC, the diffusion distance-based quantity ρ_t is replaced with a corresponding Euclidean distance-based quantity:

$$\rho^{\text{Euc}}(x) = \begin{cases} \min\limits_{y \in X} \{\|x - y\|_2 \mid p(y) \geq p(x)\}, & x \neq \arg\max\limits_{y \in X} p(y), \\ \max\limits_{y \in X} \|x - y\|_2, & x = \arg\max\limits_{y \in X} p(y). \end{cases}$$

Moreover, the modes are estimated using only $\rho^{\text{Euc}}(x)$, rather than $\mathcal{D}_t(x) = p(x)\rho_t(x)$ as proposed in the LUND algorithm. As in LUND, FSFDPC iteratively assigns points the same label as their nearest Euclidean neighbor (LUND uses diffusion nearest neighbor) of higher density.

The differences between LUND and FSFDPC are fundamental. Theoretical guarantees for the FDFDPC using Euclidean distances do not accommodate a rich class of distributions and the guarantees proved in this article fail when using $\rho^{\text{Euc}}(x)$ (as in Rodriguez and Laio (2014)) or $\mathcal{D}^{\text{euc}}(x) = p(x)\rho^{\text{Euc}}(x)$ for computing modes. This is because for clusters that are multimodal or supported near non-spherical sets, there is no reason for high-density regions of one cluster to be well-separated in Euclidean distance. In Section 6, we shall see FSFDPC fails for the motivating data in Section 3. Moreover, the use of the product $\mathcal{D}_t(x) = p(x)\rho_t(x)$ to determine modes gives LUND robustness to outliers that FSFDPC lacks. Indeed, outlying points may admit very large $\rho_t(x)$ or $\rho^{\text{Euc}}(x)$ values, but very small $p(x)$ values. In this sense, the density factor in $\mathcal{D}_t$ ensures that outlying points are not labeled as modes. This tension between geometry and density is highlighted in Theorem 3.5.

In addition, the LUND algorithm is able to correctly estimate the number of clusters in the data, even for nonlinear or elongated clusters, using the ratio (or decay) of $\mathcal{D}_t(x)$. A similar criterion for FSFDPC is not available for clusters that are nonlinear or elongated, due to the fact that high density regions connected by many paths in the data may be very far apart in Euclidean distance, leading heuristics based on the decay of $\rho^{\mathrm{Euc}}(x)$ or $\mathcal{D}^{\mathrm{euc}}(x)$ to fail; see Section 6. We remark that for simple, spherical data sets, using these heuristics for FSFDPC may work well; this is observed for isotropic Gaussian data in Section 6.3.

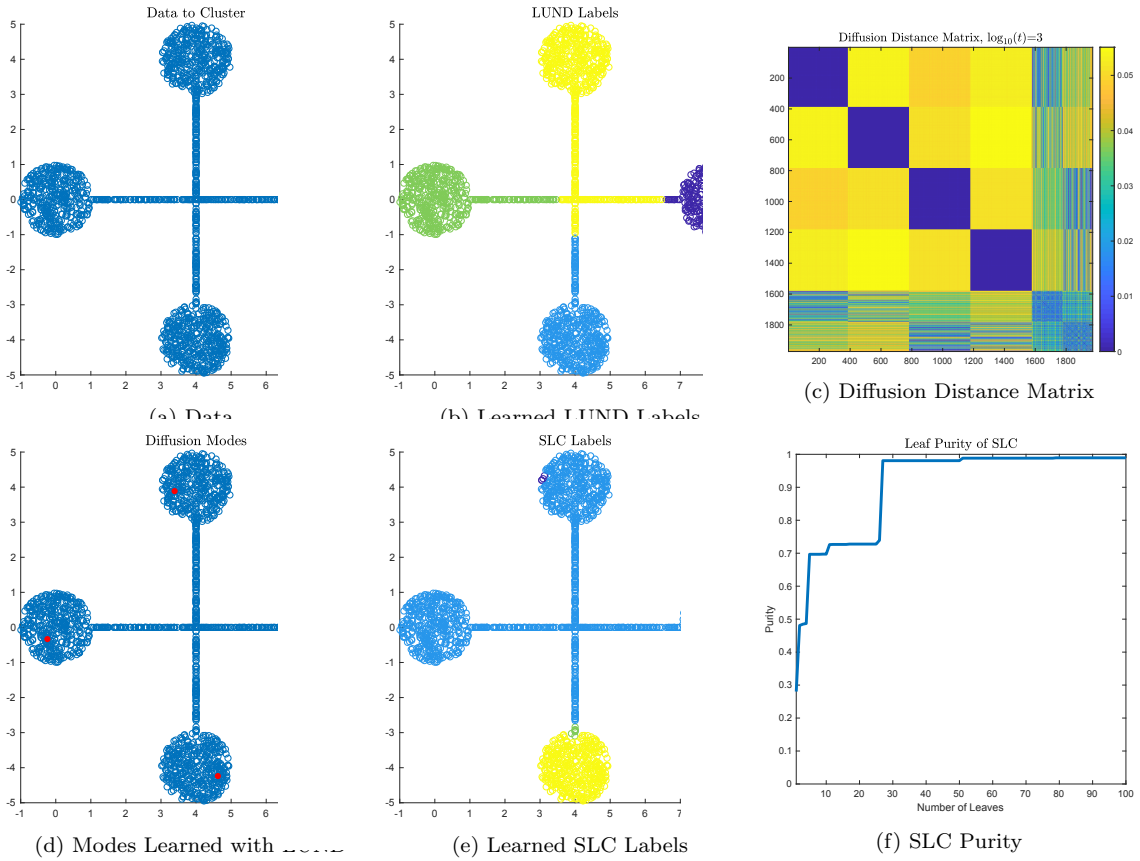
3.2.4 Comparison with Single Linkage Clustering

LUND is related to SLC in the sense that the underlying density of the data is an important determinant in the clusterings. However, LUND also incorporates geometric structure in the data when determining clusters, which can be especially powerful when density is uninformative. In Figure 4, data with constant density but significant geometric structure is analyzed. LUND succeeds in learning an accurate clustering with respect to the latent data geometry, while SLC fails. A theoretical analysis of the bottleneck phenomenon is presented in Section 4.3.

4. Analysis of Diffusion Processes on Data

In this section, we derive estimates for diffusion distances. Let $D_t^{\mathrm{in}}, D_t^{\mathrm{btw}}$ be as in (3.2). The main result of this section is to show there exists a time interval $\mathcal{T} \subset [0, \infty]$ so that $\forall t \in \mathcal{T},\ D_t^{\mathrm{btw}} > D_t^{\mathrm{in}}$, that is, for $t \in \mathcal{T}$, within cluster diffusion distance is smaller than between cluster diffusion distance. Showing that within-cluster distances are small and between-cluster distances are large is essential for any clustering problem. The benefit of using diffusion distance is its adaptability to the geometry of the data: it is possible that within cluster diffusion distance is less than between cluster diffusion distance, even in the case that the clusters are highly elongated and nonlinear. This property does not hold when points are compared with Euclidean distances or many other data-independent distances.

Compared to existing methods for analyzing diffusion distances, the proposed method does not require an analysis of the localization properties of eigenfunctions of $\mathbf{P}$, which may be challenging for small or elongated clusters (Shi et al., 2009). The proposed method has analogy to the analysis of Fokker-Planck equations of Nadler and Galun (2007), in that both approaches attempt to characterize the tension between within-cluster similarity and between-cluster separation for clustering with spectral methods. Unlike that analysis of the continuum stochastic PDE, our analysis provides explicit estimates on the behaviors of the discrete diffusion distances for finite data. Our analysis is based on the notions of *near-reducibility* and *stochastic complement*, which make precise the tension between within-cluster mixing and between-cluster transitions.



(a) Data (b) Learned LUND Labels (c) Diffusion Distance Matrix

(d) Modes Learned with LUND (e) Learned SLC Labels (f) SLC Purity

Figure 4: Geometric data with roughly constant empirical density is shown in (a). The spherical clusters are connected with thin bottlenecks. In (d), the modes learned by LUND are shown in red, indicating that one mode per cluster is learned. This is because although the density is roughly constant, the thin bottlenecks force the four circular clusters to be far apart in diffusion distance. The corresponding labels learned by LUND are in (b), showing that the cluster structure is accurately inferred. In (c), the matrix of pairwise diffusion distances is shown with diffusion time $t = 10^3$ and diffusion scaling parameter $\sigma = .5$. The four blue blocks on the diagonal indicate that within-cluster diffusion distances are small, while between cluster diffusion distances are large. The connecting bottlenecks correspond to the final rows and columns of this matrix, where the diffusion distances are less informative. In (e), labels learned from pruning the single linkage dendrogram at the third highest merge (producing four clusters) is shown, indicating this method is not appropriate for this data. Indeed, in (f), the purity of the hierarchical clustering is shown as a function of how many leaves (i.e. clusters) are used from the dendrogram. The purity remains low until ~ 25 clusters are used, indicating that the single linkage dendrogram is unable to efficiently separate the four spherical clusters until ~ 25 clusters are used. The fundamental reason for the failure of SLC on this data is the fact that the density is essentially constant and the data set is connected, which are the properties driving SLC. Unlike LUND, SLC does not incorporate geometric information, e.g. bottlenecks, which is quite discriminative for this data set.

4.1 Near Reducibility of Diffusion Processes

Let $\mathbf{P}$ be a Markov chain defined on points X satisfying the usual assumptions with unique stationary distribution $\boldsymbol{\pi}$. We will sometimes consider $\boldsymbol{\pi}$ as a function with domain X, other times as a vector with indices $\{1, \ldots, |X|\}$.

For any initial distribution $\boldsymbol{\pi}_0$, $\lim_{t\to\infty} \boldsymbol{\pi}_0 \mathbf{P}^t = \boldsymbol{\pi}$ and moreover for any choice of $\nu = \boldsymbol{\pi}_0/\boldsymbol{\pi}$, $D_t(x,y) \to 0$ uniformly as $t \to \infty$. One can quantify the rate of this convergence by estimating the convergence rate of $\mathbf{P}$ to its stationary distribution.

Definition 4.1 *For a discrete Markov chain with transition matrix* $\mathbf{P}$ *and stationary distribution* $\boldsymbol{\pi}$*, the* relative pointwise distance to $\boldsymbol{\pi}$ at time t *is* $\Delta(t) = \max_{i,j\in\{1,\ldots,n\}} |P_{ij}^t - \boldsymbol{\pi}_j|/\boldsymbol{\pi}_j$.

The decay of $\Delta(t)$ is regulated by the spectrum of $\mathbf{P}$ (Jerrum and Sinclair, 1989; Sinclair and Jerrum, 1989). Indeed, let $1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_n > -1$ be the eigenvalues of $\mathbf{P}$; note that $\lambda_2 < 1$ follows from $\mathbf{P}$ irreducible and $\lambda_n > -1$ follows from $\mathbf{P}$ aperiodic (Chung, 1997). Let $\lambda_* = \max_{i=2,\ldots,n} |\lambda_i| = \max(|\lambda_2|, |\lambda_n|)$, $\boldsymbol{\pi}_{\min} = \min_{x\in X} \boldsymbol{\pi}(x)$.

Theorem 4.2 *(Jerrum and Sinclair, 1989; Sinclair and Jerrum, 1989) Let* $\mathbf{P}$ *be the transition matrix of a Markov chain on state space* X *satisfying the usual assumptions. Then* $\Delta(t) \leq \lambda_*^t/\boldsymbol{\pi}_{\min}$.

Instead of analyzing λ_*, the *conductance* of X may be used to bound $\Delta(t)$.

Definition 4.3 *Let* $\mathcal{G}$ *be a weighted graph on* X *and let* $S \subset X$*. The* conductance of S *is* $\Phi_X(S) = \sum_{x_i\in S, x_j\in S^c} \boldsymbol{\pi}_i P_{ij} / \min\left(\sum_{x_i\in S} \boldsymbol{\pi}_i, \sum_{x_i\in S^c} \boldsymbol{\pi}_i\right)$*. The* conductance of $\mathcal{G}$ *is* $\Phi(\mathbf{P}) = \min_{S\subseteq\mathcal{G}} \Phi_X(S)$.

Methods for estimating the conductance of certain graphs include *Poincaré estimates* (Diaconis and Stroock, 1991; Diaconis and Saloff-Coste, 1993) and the method of *canonical paths* (Jerrum and Sinclair, 1989; Sinclair and Jerrum, 1989; Aldous and Fill, 2002). These approaches estimate $\Phi(\mathbf{P})$ by showing that certain simple paths may be used as surrogates for generic paths in the graph. The conductance is related to λ_2; see Chung (1997):

Theorem 4.4 *(Cheeger's Inequality) Let* $\mathcal{G}$ *be a weighted, undirected graph with transition matrix* $\mathbf{P}$*. Then the second eigenvalue* λ_2 *of* $\mathbf{P}$ *satisfies* $\Phi(\mathbf{P})^2/2 \leq 1 - \lambda_2 \leq 2\Phi(\mathbf{P})$.

Combining Theorem 4.2 and Cheeger's inequality relates $\Delta(t)$ to $\Phi(\mathbf{P})$.

Theorem 4.5 *(Jerrum and Sinclair, 1989; Sinclair and Jerrum, 1989) Let* $\mathbf{P}$ *be the transition matrix for a Markov chain on* X *satisfying the usual assumptions. Suppose* $P_{ii} \geq \frac{1}{2},\ \forall i = 1,\ldots,n$*. Then* $\Delta(t) \leq (1 - \frac{1}{2}\Phi(\mathbf{P})^2)^t/\boldsymbol{\pi}_{\min}$.

Note that any Markov chain can be made to satisfy $P_{ii} \geq \frac{1}{2}, \forall i = 1, \ldots, n$, simply by replacing $\mathbf{P}$ with $\frac{1}{2}(\mathbf{P} + \mathbf{I})$. This keeps the same stationary distribution and reduces the conductance by a factor of $\frac{1}{2}$.

Whether Theorem 4.2 or 4.5 is used, the convergence of $\mathbf{P}$ towards its stationary distribution is exponential, with rate determined by λ_* or $\Phi(\mathbf{P})$, that is, to how close to being reducible the chain is. This yields estimates on diffusion distances; indeed, for $x, y \in X$ and any initial distribution $\boldsymbol{\pi}_0$,

$$\begin{aligned} D_t(x,y) =& ||p_t(x,\cdot) - p_t(y,\cdot)||_{\ell^2(\nu)} \leq ||p_t(x,\cdot) - \boldsymbol{\pi}(\cdot)||_{\ell^2(\nu)} + ||p_t(y,\cdot) - \boldsymbol{\pi}(\cdot)||_{\ell^2(\nu)} \\ \leq& 2\sqrt{\sum_{u \in X} \max_{z \in X} \frac{|p_t(z,u) - \boldsymbol{\pi}(u)|^2}{\boldsymbol{\pi}(u)^2} \boldsymbol{\pi}(u)\boldsymbol{\pi}_0(u)} \\ \leq& 2\Delta(t)\sqrt{\sum_{u \in X} \boldsymbol{\pi}(u)\boldsymbol{\pi}_0(u)} \leq 2\Delta(t) \leq \frac{2(1 - \frac{1}{2}\Phi(\mathbf{P})^2)^t}{\boldsymbol{\pi}_{\min}}. \end{aligned}$$

Thus, as $t \to \infty$, $D_t \to 0$ uniformly at an exponential rate depending on the conductance of the underlying graph; a similar result holds for λ_* in place of $\Phi(\mathbf{P})$. This gives a global estimate on the diffusion distance in terms of λ_* and $\Phi(\mathbf{P})$. Note that a similar conclusion holds by analyzing (2.2), recalling that ψ_1 is constant and $\lambda_2 = \max_{i \neq 1} |\lambda_i| = \lambda_*$.

Unfortunately, a global estimate on diffusion distances may be too coarse for unsupervised clustering. To obtain the desired separation of $D_t^{\text{in}}, D_t^{\text{btw}}$, we need to study not the global mixing time, but the *mesoscopic* mixing times, corresponding to the time it takes for convergence of points in each cluster towards their mesoscopic equilibria, before reaching the global equilibrium. For this purpose, we use results from the theory of *nearly reducible Markov processes* (Simon and Ando, 1961; Meyer, 1989). Suppose the matrix $\mathbf{P}$ is irreducible; write $\mathbf{P}$, possibly after a permutation of the indices of the points, in block decomposition as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \ldots & \mathbf{P}_{1m} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \ldots & \mathbf{P}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{m1} & \mathbf{P}_{m2} & \ldots & \mathbf{P}_{mm} \end{bmatrix}, \tag{4.6}$$

where each $\mathbf{P}_{ii}$ is square and $m \leq n$. Let I_i be the indices of the points corresponding to $\mathbf{P}_{ii}$. Recall that if the graph corresponding to $\mathbf{P}$ is disconnected, then $\mathbf{P}$ is a *reducible* Markov chain. Recall that $\|\mathbf{A}\|_\infty = \max_i \sum_j |A_{ij}|$ is the maximal row sum of $\mathbf{A} = (A_{ij})$. Suppose that $\|\mathbf{P}_{ij}\|_\infty$ is small but nonzero for $i \neq j$, that is, most of the interactions for points in I_i are contained within $\mathbf{P}_{ii}$. This suggests diffusion on the blocks $\mathbf{P}_{ii}$ have dynamics that converge to their own, mesoscopic equilibria before the entire chain converges to a global equilibrium, depending on the weakness of connection between blocks. Interpreting the support sets I_i as corresponding to the clusters of X, this suggests there will be a time range for which points within

each cluster are close in diffusion distance but far in diffusion distance from points in other clusters; such a state corresponds to a mesoscopic equilibrium. To make this precise, consider the notion of *stochastic complement*.

Definition 4.7 *Let* $\mathbf{P}$ *be an* $n \times n$ *irreducible Markov matrix partitioned into square block matrices as in (4.6). For a given index* $i \in \{1, \dots, m\}$*, let* $\mathbf{P}_i$ *denote the principal block submatrix generated by deleting the* i^{th} *row and* i^{th} *column of blocks from (4.6), and let* $\mathbf{P}_{*i} = \left[\mathbf{P}_{1i}\mathbf{P}_{2i}\dots\mathbf{P}_{i-1,i}\mathbf{P}_{i+1,i}\dots\mathbf{P}_{mi}\right]^T$ *and* $\mathbf{P}_{i*} = \left[\mathbf{P}_{i1}\ \mathbf{P}_{i2}\ \dots\ \mathbf{P}_{i,i-1}\ \mathbf{P}_{i,i+1}\ \dots\ \mathbf{P}_{im}\right]$. *The* stochastic complement *of* $\mathbf{P}_{ii}$ *is the matrix* $\mathbf{S}_{ii} = \mathbf{P}_{ii} + \mathbf{P}_{i*}(\mathbf{I} - \mathbf{P}_i)^{-1}\mathbf{P}_{*i}$.

One can interpret the stochastic complement $\mathbf{S}_{ii}$ as the transition matrix for a reduced Markov chain obtained from the original chain, but in which transitions into or out of I_i are masked. More precisely, in the reduced chain $\mathbf{S}_{ii}$, a transition is either direct in $\mathbf{P}_{ii}$ or indirect by moving first through points outside of $\mathbf{P}_{ii}$, then back into $\mathbf{P}_{ii}$ at some future time. Indeed, the term $\mathbf{P}_{i*}(\mathbf{I} - \mathbf{P}_i)^{-1}\mathbf{P}_{*i}$ in the definition of $\mathbf{S}_{ii}$ accounts for leaving I_i (the factor $\mathbf{P}_{i*}$), traveling for some time in I_i^c (the factor $(\mathbf{I} - \mathbf{P}_i)^{-1}$), then re-entering I_i (the factor $\mathbf{P}_{*i}$). Note that the factor $(\mathbf{I} - \mathbf{P}_i)^{-1}$ may be expanded in Neumann sum as $(\mathbf{I} - \mathbf{P}_i)^{-1} = \sum_{t=0}^{\infty} \mathbf{P}_i^t$, showing that it accounts for exiting from I_i and returning to it after an arbitrary number of steps outside of it.
The notion of stochastic complement quantifies the interplay between the mesoscopic and global equilibria of $\mathbf{P}$. We say $\mathbf{P}$ is *primitive* if it is non-negative, irreducible and aperiodic. The following theorem indicates how $\mathbf{P}$ may be analyzed when it is derived from cluster data $\{X_k\}_{k=1}^K$ sampled according to (3.1); a proof appears in the Appendix for completeness. This result, which produces estimates related to the diffusion operator $\mathbf{P}$ in the ℓ^1 norm, is used to prove results on diffusion distances, which are defined in an ℓ^2 sense, partially in order to take advantage of spectral decompositions for fast computations. This discrepancy will be discussed and controlled in Section 4.2.

Theorem 4.8 *(Meyer, 1989) Let* $\mathbf{P}$ *be an* $n \times n$ *irreducible row-stochastic matrix partitioned into* K^2 *square block matrices, and let* $\mathbf{S}$ *be the reducible row-stochastic matrix consisting of the stochastic complements of the diagonal blocks of* $\mathbf{P}$*:*

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \dots & \mathbf{P}_{1K} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \dots & \mathbf{P}_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{P}_{K1} & \mathbf{P}_{K2} & \dots & \mathbf{P}_{KK} \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} \mathbf{S}_{11} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{S}_{22} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{S}_{KK} \end{bmatrix}.$$

Suppose each $\mathbf{S}_{ii}$ *is primitive, so that the eigenvalues of* $\mathbf{S}$ *satisfy* $\lambda_1 = \lambda_2 = \dots = \lambda_K = 1 > \lambda_{K+1} \geq \lambda_{K+2} \geq \dots > -1$*. Let* $\mathbf{Z}$ *diagonalize* $\mathbf{S}$*, and let*

$$\mathbf{S}^\infty = \lim_{t\to\infty} \mathbf{S}^t = \begin{bmatrix} \mathbf{1}\boldsymbol{\pi}^1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}\boldsymbol{\pi}^2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{1}\boldsymbol{\pi}^K \end{bmatrix},$$

where $\boldsymbol{\pi}^i$ *is the stationary distribution for* $\mathbf{S}_{ii}$*. Then* $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty \leq t\delta + \kappa|\lambda_{K+1}|^t$*, where* $\delta = 2\max_i \|\mathbf{P}_{i*}\|_\infty$ *and* $\kappa = \|\mathbf{Z}\|_\infty\|\mathbf{Z}^{-1}\|_\infty$*. Moreover, for any initial distribution* $\boldsymbol{\pi}_0$ *and* $\mathbf{s} = \lim_{t\to\infty} \boldsymbol{\pi}_0\mathbf{S}^t = \boldsymbol{\pi}_0\mathbf{S}^\infty$*,* $\|\boldsymbol{\pi}_0\mathbf{P}^t - \mathbf{s}\|_1 \leq t\delta + \kappa|\lambda_{K+1}|^t$*.*

Note that this result does not require the Markov chain to be reversible, and hence applies to diffusion processes defined on *directed* graphs. The assumption that $\mathbf{S}$ is diagonalizable is not strictly necessary, and similar estimates hold more generally (Meyer, 1989).
The estimate $t\delta + \kappa|\lambda_{K+1}|^t$ consists of two terms. The $t\delta$ term corresponds to $\|\mathbf{P}^t - \mathbf{S}^t\|_\infty$, which accounts for the approximation of $\mathbf{P}^t$ by the reducible Markov chain $\mathbf{S}^t$. In the context of clustering, this term accounts for the between-cluster connections in $\mathbf{P}$. The term $\kappa|\lambda_{K+1}|^t$ corresponds to $\|\mathbf{S}^t - \mathbf{S}^\infty\|_\infty$, which accounts for propensity of mixing within a cluster. In the clustering context, this term quantifies the within-cluster distances.
It follows from Theorem 4.8 that, given ϵ sufficiently large, there is a range of t for which the dynamics of $\mathbf{P}^t$ are ϵ-close to the dynamics of the reducible, low-rank Markov chain $\mathbf{S}^\infty$.

Corollary 4.9 *Let* $\lambda_{K+1}, \delta, \kappa$ *be as in Theorem 4.8. Suppose that for some* $\epsilon > 0$*,* $\ln\left(\frac{2\kappa}{\epsilon}\right)/\ln\left(\frac{1}{|\lambda_{K+1}|}\right) < t < \frac{\epsilon}{2\delta}$*. Then* $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty < \epsilon$*, and for every initial distribution* $\boldsymbol{\pi}_0$*,* $\|\boldsymbol{\pi}_0\mathbf{P}^t - \mathbf{s}\|_1 < \epsilon$*.*

In contrast with t, the values $\lambda_{K+1}, \delta, \kappa$ may be understood as fixed geometric parameters of the data set which determine the range of times t at which mesoscopic equilibria are reached. More precisely, as $n \to \infty$, δ, κ converge to natural continuous quantities independent of n, and Garcia Trillos et al. (2018) proved that as $n \to \infty$, there is a natural scaling for $\sigma \to 0^+$ in which the (random) empirical eigenvalues of $\mathbf{P}$ converge in a precise sense to the (deterministic) eigenvalues of a corresponding continuous operator defined on the support of μ as in (3.1). Thus, the parameters of Theorem 4.9 may be understood as random fluctuations of geometrically intrinsic quantities depending on μ. In the context of the proposed data model, these quantities may be interpreted as follows:

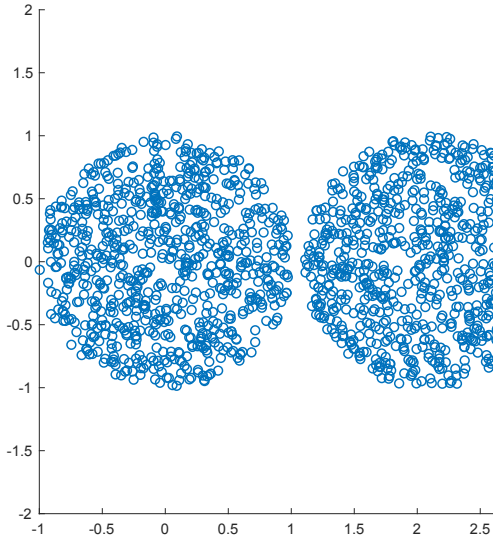
- λ_{K+1} is the largest eigenvalue of $\mathbf{S}$ not equal to 1. Since $\mathbf{S}$ is block diagonal and each $\mathbf{S}_{kk}$ is primitive, it follows that $\lambda_{K+1} = \max_{k=1,\dots,K} \lambda_2(\mathbf{S}_{kk})$. As discussed above, $\{\lambda_2(\mathbf{S}_{kk})\}_{k=1}^K$ is related to the conductance $\Phi(\mathbf{S}_{kk})$ and the mixing time of the random walk restricted to $\mathbf{S}_{kk}$. If the entries of $\mathbf{S}_{kk}$ are very close to the entries of $\mathbf{P}_{kk}$, then a perturbative argument yields $\lambda_2(\mathbf{S}_{kk}) \approx \lambda_2(\mathbf{P}_{kk})$.

- The quantity $\delta = 2\max_{k=1,\dots,K} \|\mathbf{P}_{k*}\|_\infty$ is controlled by the largest interaction between clusters. If the separation between the $\{X_k\}_{k=1}^K$ is large enough, δ will be small. To get a sense of this parameter, let C_1, C_2 be clusters with $n/2$ points uniformly sampled from the balls $B_1(0,0), B_1(2+\eta, 0) \subset \mathbb{R}^2$. Then, for

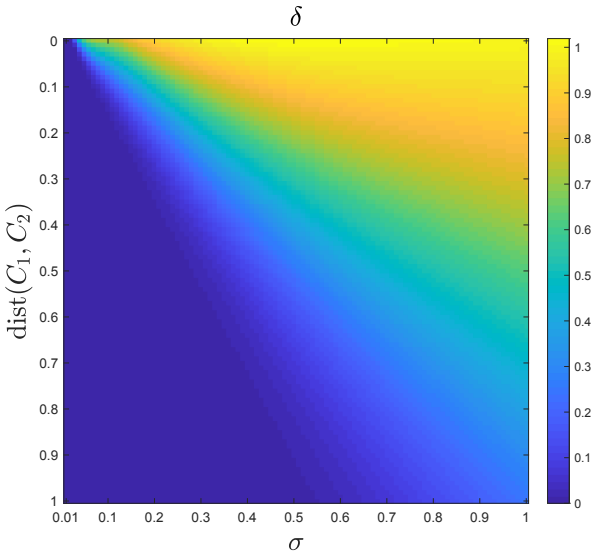
n sufficiently large, $\mathrm{dist}(C_1, C_2) = \min_{x_1 \in C_1, x_2 \in C_2} \|x_1 - x_2\|_2 \approx \eta$. Note that the point in $B_1(0,0)$ nearest $B_1(2+\eta, 0)$ is $(1,0)$. Then modulo the variance from the random sampling,

$$\delta \approx \frac{2\int_{B_1(2+\eta,0)} \exp(-\|(x,y)-(1,0)\|_2^2/\sigma^2)dxdy}{\int_{B_1(2+\eta,0)} \exp(-\|(x,y)-(1,0)\|_2^2/\sigma^2)dxdy + \int_{B_1(0,0)} \exp(-\|(x,y)-(1,0)\|_2^2/\sigma^2)dxdy} \leq \frac{2\exp(-\eta^2/\sigma^2)}{\exp(-(\eta^2+4\eta)/\sigma^2)+1}. \tag{4.10}$$

Figure 5 illustrates empirically how δ depends on σ and $\mathrm{dist}(C_1, C_2)$ for such



(a) Uniform, spherical clusters C_1, C_2 $\mathrm{dist}(C_1, C_2) \approx .1$



(b) δ as a function of $\mathrm{dist}(C_1, C_2)$ and σ.

Figure 5: In (a), data drawn uniformly at random from the union of two balls in $\mathbb{R}^2$ is shown. In (b), it is shown that for such equal sized, spherical, constant density clusters, the separation parameter exhibits rapid decay as $\sigma \to 0^+$, for fixed separation. The experiments confirm that the decay of δ is essentially logistic in $-(\eta^2/\sigma^2)$, as estimated in (4.10).

- The quantity $\kappa = \|\mathbf{Z}\|_\infty \|\mathbf{Z}^{-1}\|_\infty$, with $\mathbf{Z} = (\phi_1 | \dots | \phi_n)$, is a measure of the condition number of diagonalizing $\mathbf{S}$. If $\mathbf{Z}, \mathbf{Z}^{-1}$ are orthogonal matrices, then each row of $\mathbf{Z}, \mathbf{Z}^{-1}$ have ℓ^2 norm 1, hence $\kappa \leq n$. We remark that κ is bounded independently of n in the case that all the data live on a common manifold, using convergence of heat kernels and low-frequency eigenfunctions together with heat kernel estimates on manifolds. In the clustering setting, if each cluster is a manifold, similar results would hold in this case, albeit this analysis is a topic of ongoing research.

4.2 Diffusion Distance Estimates

Returning to the proposed data model $X = \bigcup_{k=1}^{K} X_k \sim \mu$ as per (3.1), let $\mathbf{P}$ be a corresponding Markov chain on X satisfying the usual assumptions. We estimate the dependence of diffusion distances on the parameters $\delta, \lambda_{K+1}, \kappa$ above. We also introduce a *balance quantity* that quantifies the difference between the ℓ^1 norm (the setting of Theorem 4.8) and the ℓ^2 norm (the setting of diffusion distances). Throughout this section, let $p_t(x_i, x_j) = \mathbf{P}^t_{ij},\ \mathbf{s}^\infty(x_i, x_j) = \mathbf{S}^\infty_{ij}$.

Definition 4.11 *Let $\mathbf{P}, \mathbf{S}^\infty \in \mathbb{R}^{n\times n}$ be as in Theorem 4.8. Define*

$$\gamma(t) = \max_{x\in X}\left(1 - \frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(x,u) - \mathbf{s}^\infty(x,u)|}{\|p_t(x,\cdot) - \mathbf{s}^\infty(x,\cdot)\|_2} - \frac{1}{\sqrt{n}}\right|^2\right)^{-1}.$$

Botelho-Andrade et al. (2019) show that for any vector $v \in \mathbb{R}^n$, $\|v\|_2 = \frac{c_v}{\sqrt{n}}\|v\|_1$, where

$$c_v = \left(1 - \frac{1}{2}\sum_{i=1}^{n}\left|\frac{|v_i|}{\|v\|_2} - \frac{1}{\sqrt{n}}\right|^2\right)^{-1}.$$

In this sense, $\gamma(t)$ measures how the ℓ^1 norm differs from the ℓ^2 norm across all rows of $\mathbf{P}^t - \mathbf{S}^\infty$. In particular, when each row of $\mathbf{P}^t - \mathbf{S}^\infty$ is close to uniform, $\gamma(t)$ is close to 1; when some row of $\mathbf{P}^t - \mathbf{S}^\infty$ concentrates all its mass around one index, then $\gamma(t) = \sqrt{n}$. Note that $1 \le \gamma(t) \le \sqrt{n}$ for all t.

Theorem 4.12 *Let $X = \bigcup_{k=1}^{K} X_k$ and let $\mathbf{P}$ be a corresponding Markov transition matrix on X. Let $\delta, \lambda_{K+1}, \kappa, \mathbf{S}^\infty$ be as in Theorem 4.8. Let D_t be the diffusion distance associated to $\mathbf{P}$ and counting measure ν. If t, ϵ satisfy*

$$\frac{\ln\left(\frac{2\kappa}{\epsilon}\right)}{\ln\left(\frac{1}{\lambda_{K+1}}\right)} < t < \frac{\epsilon}{2\delta},$$

then

(a) $D_t^{in} \le 2\dfrac{\epsilon}{\sqrt{n}}\gamma(t)$.

(b) $D_t^{btw} \ge 2\min_{y\in X}\|\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)} - 2\dfrac{\epsilon}{\sqrt{n}}\gamma(t)$.

Proof By Corollary 4.9, $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty < \epsilon$, that is, $\max_{x\in X}\sum_{u\in X}|p_t(x,u) - \mathbf{s}^\infty(x,u)|\nu(u) < \epsilon$. To see (a), let k be arbitrary and let $x, y \in X_k$. Then:

$$\begin{aligned}&||p_t(x,\cdot) - p_t(y,\cdot)||_{\ell^2(\nu)}\\ \le&||p_t(x,\cdot) - \mathbf{s}^\infty(x,\cdot)||_{\ell^2(\nu)} + ||p_t(y,\cdot) - \mathbf{s}^\infty(y,\cdot)||_{\ell^2(\nu)} + ||\mathbf{s}^\infty(y,\cdot) - \mathbf{s}^\infty(x,\cdot)||_{\ell^2(\nu)}\end{aligned}$$

$$
\begin{aligned}
=&\frac{1}{\sqrt{n}}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(x,u)-\mathbf{s}^\infty(x,u)|}{\|p_t(x,\cdot)-\mathbf{s}^\infty(x,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1}||p_t(x,\cdot)-\mathbf{s}^\infty(x,\cdot)||_{\ell^1(\nu)}\\
+&\frac{1}{\sqrt{n}}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(y,u)-\mathbf{s}^\infty(y,u)|}{\|p_t(y,\cdot)-\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1}||p_t(y,\cdot)-\mathbf{s}^\infty(y,\cdot)||_{\ell^1(\nu)}\\
+&||\mathbf{s}^\infty(y,\cdot)-\mathbf{s}^\infty(x,\cdot)||_{\ell^2(\nu)}\\
\le&\frac{2\epsilon}{\sqrt{n}}\max_{x\in X}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(x,u)-\mathbf{s}^\infty(x,u)|}{\|p_t(x,\cdot)-\mathbf{s}^\infty(x,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1}+||\mathbf{s}^\infty(y,\cdot)-\mathbf{s}^\infty(x,\cdot)||_{\ell^2(\nu)},
\end{aligned}
$$

where t satisfies $\ln\left(\frac{2\kappa}{\epsilon}\right)/\ln\left(\frac{1}{\lambda_{K+1}}\right) < t < \epsilon/(2\delta)$. The line relating the norm in $\ell^1(\nu)$ and $\ell^2(\nu)$ follows from Theorem 1 in Botelho-Andrade et al. (2019). Note that $\mathbf{S}^\infty$ has constant columns on each cluster, and in particular for $x, y \in X_k$, $\mathbf{s}^\infty(x,u) = \mathbf{s}^\infty(y,u) = \boldsymbol{\pi}^k(u)$ for all $u \in X$, so that $||\mathbf{s}^\infty(y,\cdot)-\mathbf{s}^\infty(x,\cdot)||_{\ell^2(\nu)} = 0$. Statement (a) follows.

To see (b), suppose that $x \in X_k, y \in X_\ell,\ k \neq \ell$. Then

$$
\begin{aligned}
&\|p_t(x,\cdot)-p_t(y,\cdot)\|_{\ell^2(\nu)}\\
=&\|p_t(x,\cdot)-\mathbf{s}^\infty(x,\cdot)+\mathbf{s}^\infty(x,\cdot)-\mathbf{s}^\infty(y,\cdot)+\mathbf{s}^\infty(y,\cdot)-p_t(y,\cdot)\|_{\ell^2(\nu)}\\
\ge&\|\mathbf{s}^\infty(x,\cdot)-\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}-\|p_t(x,\cdot)-\mathbf{s}^\infty(x,\cdot)\|_{\ell^2(\nu)}-\|p_t(y,\cdot)-\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}\\
=&\|\mathbf{s}^\infty(x,\cdot)-\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}\\
&-\frac{1}{\sqrt{n}}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(x,u)-\mathbf{s}^\infty(x,u)|}{\|p_t(x,\cdot)-\mathbf{s}^\infty(x,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1}||p_t(x,\cdot)-\mathbf{s}^\infty(x,\cdot)||_{\ell^1(\nu)}\\
&-\frac{1}{\sqrt{n}}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(y,u)-\mathbf{s}^\infty(y,u)|}{\|p_t(y,\cdot)-\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1}||p_t(y,\cdot)-\mathbf{s}^\infty(y,\cdot)||_{\ell^1(\nu)}\\
\ge&\|\mathbf{s}^\infty(x,\cdot)-\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}-\frac{\epsilon}{\sqrt{n}}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(x,u)-\mathbf{s}^\infty(x,u)|}{\|p_t(x,\cdot)-\mathbf{s}^\infty(x,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1}\\
&-\frac{\epsilon}{\sqrt{n}}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(y,u)-\mathbf{s}^\infty(y,u)|}{\|p_t(y,\cdot)-\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1}\\
\ge&\|\mathbf{s}^\infty(x,\cdot)-\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}-2\frac{\epsilon}{\sqrt{n}}\max_{z\in X_\ell\cup X_k}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(z,u)-\mathbf{s}^\infty(z,u)|}{\|p_t(z,\cdot)-\mathbf{s}^\infty(z,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1}\\
\ge&2\min_{w\in X_k\cup X_\ell}\|\mathbf{s}^\infty(w,\cdot)\|_{\ell^2(\nu)}-2\frac{\epsilon}{\sqrt{n}}\max_{z\in X_\ell\cup X_k}\left(1-\frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(z,u)-\mathbf{s}^\infty(z,u)|}{\|p_t(z,\cdot)-\mathbf{s}^\infty(z,\cdot)\|_{\ell^2(\nu)}}-\frac{1}{\sqrt{n}}\right|^2\right)^{-1},
\end{aligned}
$$

where in the last step, to lower bound the first term we used that $\mathbf{s}^\infty(y,\cdot) = \boldsymbol{\pi}^l(\cdot)$, $\mathbf{s}^\infty(x,\cdot) = \boldsymbol{\pi}^k(\cdot)$, and recalled that since $k \neq l$ the supports of $\mathbf{s}^\infty(x,\cdot)$ and $\mathbf{s}^\infty(y,\cdot)$

are disjoint. Minimizing this lower bound over all clusters X_k, X_ℓ yields the desired result. ■

Heuristically, if ϵ is small and the reduced equilibrium distribution $\mathbf{s}^\infty$ is roughly constant on each cluster, there will be a range of t for which $D_t^{\text{in}} \ll D_t^{\text{btw}}$. The notion of $\mathbf{s}^\infty$ being roughly constant on each cluster is equivalent to nodes in the same cluster having roughly constant degree. These theoretical estimates are compared to empirical bounds computed numerically in Section 6.
If $\mathbf{P}$ is very close to $\mathbf{S}$ in Frobenius norm, then $p_t(x,y)$ is very close to $\mathbf{s}^\infty(x,y)$ and ϵ may be taken close to 0. In particular, for the ideal case $\epsilon = 0$, the estimates of Theorem 4.12 reduce to

$$D_t^{\text{in}} = 0 \qquad , \qquad D_t^{\text{btw}} \geq 2 \min_{y \in X} \|\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}. \tag{4.13}$$

One can define a natural notion of diffusion distance between disjoint clusters in a reducible Markov chain as the sum of the ℓ^2 norms of their respective stationary distribution, which agrees with both the definition of diffusion distances upon taking the limit $t \to +\infty$ and with the lower bound (b) in Theorem 4.12 when $\epsilon \to 0^+$. Hence, while the estimates in the proof of Theorem 4.12 may not be optimal, they are quite natural for $\epsilon \to 0^+$.
Away from the asymptotic regime $\epsilon \to 0^+$, the estimates of Theorem 4.12 may be further simplified by placing additional assumptions on the data. Indeed, if the equilibrium distributions in $\mathbf{S}^\infty$ are balanced and uniform, the following result holds:

Corollary 4.14 *Suppose that $\mathbf{s}^\infty$ is uniform on each X_k, and the cardinality of each X_k is n/K. Then for any t, ϵ satisfying $\ln\left(\frac{2\kappa}{\epsilon}\right)/\ln\left(\frac{1}{\lambda_{K+1}}\right) < t < \frac{\epsilon}{2\delta}$,*

$$D_t^{in} \leq \frac{2}{\sqrt{n}} \epsilon \gamma(t) \qquad , \qquad D_t^{btw} \geq \frac{2}{\sqrt{n}} \left(\sqrt{K} - \epsilon\gamma(t)\right).$$

Proof If $\mathbf{S}^\infty$ has constant rows on each cluster (i.e. the stationary distribution on each cluster of the reduced Markov chain is uniform), and the clusters are of constant size n/K, then $2\min_{y\in X} \|\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)} = 2\sqrt{K/n}$. Then Theorem 4.12 yields

$$\begin{aligned} D_t^{\text{btw}} \geq & 2 \min_{y \in X} \|\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)} - 2\frac{\epsilon}{\sqrt{n}}\gamma(t) \\ = & 2\sqrt{\frac{K}{n}} - 2\frac{\epsilon}{\sqrt{n}}\gamma(t) \\ = & \frac{2}{\sqrt{n}}\left(\sqrt{K} - \epsilon\gamma(t)\right). \end{aligned}$$

■

In particular, if $\epsilon \ll \sqrt{K}/(2\gamma(t))$, within cluster distances will be small since $D_t^{\text{in}} \ll \sqrt{K/n}$, and also there will be clear separation between clusters since $D_t^{\text{btw}} = \Omega(\sqrt{K/n})$. Note that when $\mathbf{P}^t - \mathbf{S}^\infty$ is balanced, $\gamma(t)$ is $O(1)$ with respect to n, so that the assumption on ϵ is independent of n.
We remark that in general if $\mathbf{S}^\infty$ and $\mathbf{P}^t - \mathbf{S}^\infty$ are balanced in the sense of having approximately uniform rows, then $\min_{y\in X} \|\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)}$ scales like $1/\sqrt{n}$ and $\gamma(t) = O(1)$, respectively. In this (generic) case, the bounds of Theorem 4.12 are on the order $1/\sqrt{n}$. This is natural, since diffusion distances are computed in the ℓ^2 sense, while $\mathbf{P}^t$ and $\mathbf{S}^\infty$ have rows with ℓ^1 norm equal to 1.

4.3 Relaxing Separation Between Clusters

The analysis of diffusion distances in Section 4.2 depends on δ, a parameter characterizing separation between clusters. When δ is large, the estimates on D_t^{in} and D_t^{btw} from Theorem 4.12 are poor. However, δ is defined as the *maximum over all points* of the probability of transitioning to a new cluster in one time step. The formulation in terms of the maximum is convenient for analysis, but is pessimistic in the case of clusters cores that are separated by bottlenecks or low density noise regions. Indeed, for these data, δ may be large—there are some points quite close to the cluster boundaries—but the probability of transition between clusters cores is low.
The conservative estimates based on δ may be improved by considering some data points not as part of a cluster, but rather as noise or transition points. Let $X = \bigcup_{k=1}^K X_k \cup \mathcal{N}$ be a decomposition of X into cluster points ($\bigcup_{k=1}^K X_k$) and noise points ($\mathcal{N}$). Decompose $\mathbf{P}$ as

$$\mathbf{P} = \begin{bmatrix} \mathbf{P}_{11} & \mathbf{P}_{12} & \dots & \mathbf{P}_{1K} & \tilde{\mathbf{P}}_{1,\mathcal{N}} \\ \mathbf{P}_{21} & \mathbf{P}_{22} & \dots & \mathbf{P}_{2K} & \tilde{\mathbf{P}}_{2,\mathcal{N}} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{P}_{K1} & \mathbf{P}_{K2} & \dots & \mathbf{P}_{KK} & \tilde{\mathbf{P}}_{K,\mathcal{N}} \\ \tilde{\mathbf{P}}_{\mathcal{N},1} & \tilde{\mathbf{P}}_{\mathcal{N},2} & \dots & \tilde{\mathbf{P}}_{\mathcal{N},K} & \tilde{\mathbf{P}}_{\mathcal{N},\mathcal{N}} \end{bmatrix} \tag{4.15}$$

in a manner analogous to (4.6). Interpreting the final block rows and columns as noise or bottleneck points that are not part of clusters, Theorem 4.12 may be generalized to allow for clusters that are not well-separated, by accounting for the geometric properties of noise and bottlenecks which do not admit transitions between clusters over short time scales. This more generalized analysis does not make use of the reduced Markov chain $\mathbf{S}$, and some slightly new notation is required.

Definition 4.16 *For* $\mathbf{P}$ *as in (4.15) define*

$$\gamma_{min}(t) = \min_{x,y\in X, x\neq y} \left(1 - \frac{1}{2}\sum_{u\in X} \left| \frac{|p_t(x,u) - p_t(y,u)|}{\|p_t(x,\cdot) - p_t(y,\cdot)\|_2} - \frac{1}{\sqrt{n}} \right|^2 \right)^{-1},$$

$$\gamma_{max}(t) = \max_{x,y\in X}\left(1 - \frac{1}{2}\sum_{u\in X}\left|\frac{|p_t(x,u)-p_t(y,u)|}{\|p_t(x,\cdot)-p_t(y,\cdot)\|_2} - \frac{1}{\sqrt{n}}\right|^2\right)^{-1}.$$

The following result is more general than Theorem 4.12, in that it allows for bottlenecks in the data, but is not interpretable in terms of the near reducibility of $\mathbf{P}$. Let $D_t^{\text{in}}, D_t^{\text{btw}}$ be as in (3.2); note that these quantities ignore $\mathcal{N}$ and only consider distances within and between the clusters $\{X_k\}_{k=1}^K$.

Theorem 4.17 *Let $X = \bigcup_{k=1}^K X_k \cup \mathcal{N}$ and let $\mathbf{P}$ be a corresponding Markov transition matrix on X, decomposed as in (4.15), and let ν be the counting measure. Let*

$$\begin{aligned}
\delta(t) =& 2\max_{1\le k\le K}\|(\mathbf{P}^t)_{k*}\|_\infty, \text{ where } \mathbf{P}_{k*} = [\mathbf{P}_{k1}\mathbf{P}_{k2}\dots\mathbf{P}_{k,k-1}\mathbf{P}_{k,k+1}\dots\mathbf{P}_{kK}],\\
\alpha(t) =& \max_{1\le k\le K}\max_{x,y\in X_k}\sum_{u\in X_k}|p_t(x,u)-p_t(y,u)|,\\
\beta(t) =& 2\min_{1\le k\le K}\min_{x\in X_k}\sum_{u\in X_k}|p_t(x,u)|,\\
\zeta(t) =& \max_{1\le k\le K}\max_{x,y\in X_k}\sum_{u\in \mathcal{N}}|p_t(x,u)-p_t(y,u)|.
\end{aligned}$$

Then:

(a) $D_t^{in} \le \gamma_{max}(t)(\alpha(t)+\delta(t)+\zeta(t))$.

(b) $D_t^{btw} \ge \gamma_{min}(t)(\beta(t)-\delta(t))$.

Proof To see (a), suppose $x, y \in X_k$. Then

$$\begin{aligned}
&\|p_t(x,\cdot)-p_t(y,\cdot)\|_{\ell^2(\nu)}\\
\le&\gamma_{\max}(t)\|p_t(x,\cdot)-p_t(y,\cdot)\|_{\ell^1(\nu)}\\
=&\gamma_{\max}(t)\left(\sum_{u\in X_k}|p_t(x,u)-p_t(y,u)| + \sum_{\ell\ne k}\sum_{u\in X_\ell}|p_t(x,u)-p_t(y,u)| + \sum_{u\in\mathcal{N}}|p_t(x,u)-p_t(y,u)|\right)\\
\le&\gamma_{\max}(t)(\alpha(t)+\delta(t)+\zeta(t)).
\end{aligned}$$

On the other hand, suppose $x \in X_k, y \in X_\ell, k \ne \ell$. Then

$$\begin{aligned}
\|p_t(x,\cdot)-p_t(y,\cdot)\|_{\ell^2(\nu)} \ge&\gamma_{\min}(t)\|p_t(x,\cdot)-p_t(y,\cdot)\|_{\ell^1(\nu)}\\
\ge&\gamma_{\min}(t)\left(\sum_{u\in X_k}|p_t(x,u)-p_t(y,u)| + \sum_{u\in X_\ell}|p_t(x,u)-p_t(y,u)|\right)\\
\ge&\gamma_{\min}(t)\left(\sum_{u\in X_k}|p_t(x,u)| - |p_t(y,u)| + \sum_{u\in X_\ell}|p_t(y,u)| - |p_t(x,u)|\right)
\end{aligned}$$

$$\geq \gamma_{\min}(t)(\beta(t) - \delta(t)).$$

■

The quantities $\delta(t), \alpha(t), \beta(t)$ in Theorem 4.17 are analogous to those in Theorem 4.12, albeit without the aid of the analysis in terms of $\mathbf{S}^\infty$. The quantity $\zeta(t)$ pertains to the noise $\mathcal{N}$, and hence has no analogy in Theorem 4.12.

- The quantity $\delta(t)$ is an upper bound on the probability mass that transitions between clusters at time t. It does not consider the mass transfer from clusters to the noise region. In Theorem 4.8, when the clusters are well-separated and $\mathcal{N} = \varnothing$, $\delta(t) \leq \delta(1)t = \delta t$ is a worst-case, though intuitive, upper bound. Generally, $\delta(t)$ may depend on time nonlinearly and in particular it may stay small for a long time, before the global equilibrium time of the underlying Markov chain $\mathbf{P}$ is approached, at which point it grows essentially linearly in t before stabilizing around the global mixing time. This occurs, for example, in data with bottlenecks between clusters. We remark that considering $\delta(t)$ in these more general terms is possible even in the noiseless case of Theorem 4.8. Indeed, this may allow for a better estimate than the worst-case estimate of $\delta(t) \leq \delta(1)t = \delta t$, though at some loss of simplicity.

- The quantity $\alpha(t)$ controls how close the diagonal blocks are to having constant rows, that is, how close to mesoscopic equilibria the random walker is on each individual cluster. In the case that all individual clusters are near their mesoscopic equilibria by time t, $\alpha(t)$ is close to 0. In the simpler context of Theorem 4.8, $\alpha(t)$ is governed by λ_{K+1}^t, that is, by the second largest eigenvalue of the blocks of the stochastic complement of $\mathbf{P}$. This is because the second smallest eigenvalue on each block controls the rate of convergence to local equilibrium of that block in the stochastic complement $\mathbf{S}$.

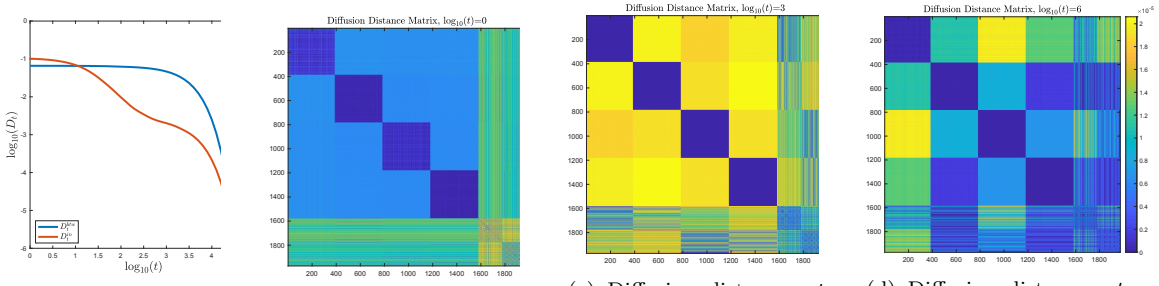
- The quantity $\beta(t)$ measures the row sums of the transition matrices localized to each cluster. It is analogous to the term involving the norm of the stationary distribution in the lower bound on D_t^{btw} in Theorem 4.12.

- The quantity $\zeta(t)$ is a measure of the impact of the noise region in the computation of diffusion distances. When the number of noise points is small relative to the number of cluster points, this term is negligible. It is also negligible when the noise region impacts a cluster in a uniform way, in particular when the cluster has reached a mesoscopic equilibrium.

Theorem 4.17 suggests that for data in which, with high probability, the clusters reach mesoscopic equilibria before the random walker exits the clusters and crosses the noise, there exists a range of parameters t such that D_t^{in} is small and D_t^{btw} is large. At this time scale, $\alpha(t)$ is small since the individual clusters are near mesoscopic equilibria,

and $\delta(t)$ is small and $\beta(t)$ large because most of the mass is still localized on the distinct clusters.

An important class of examples exhibiting this behavior is clusters connected by narrow bottlenecks, as shown for example in Figure 4 (a). In that example, consider the four balls as clusters $\{X_k\}_{k=1}^4$ and the bottlenecks as $\mathcal{N}$. The behavior of D_t^{in} and D_t^{btw} as a function of t, as well as several transition matrices for different t values are shown in Figure 6.



(a) $D_t^{\text{btw}}, D_t^{\text{in}}$ versus t. (b) Diffusion distances, $t = 1$ (c) Diffusion distances, $t = 10^3$ (d) Diffusion distances, $t = 10^6$

Figure 6: Plots illustrating that the bottleneck data in Figure 4 exhibits mesoscopic equilibria, despite the lack of any strict separation between the four clusters. Indeed, (a) shows that for a large time range, D_t^{in} is much less (note that the vertical axis is in $\log_{10}$ scale) than D_t^{btw}. This is further illustrated by the fact that for different ranges of time t, the blocks in the diffusion distance matrix corresponding to the cluster sets become essentially all 0—see (b), (c). For large time, the diffusion distances start to converge uniformly to 0, as shown in (d). Note that the last rows and columns of $\mathbf{P}$ correspond to the bottleneck points.

Theorem 4.12 does not provide useful estimates for this bottleneck data, because of the lack of strict separation between the clusters. Indeed, if bottleneck points are assigned to their nearest clusters, then δ is nearly 1, rendering Theorem 4.12 essentially useless. However, Theorem 4.17 supplies useful estimates on the diffusion distances within and between clusters (e.g. that $D_t^{\text{in}} \ll D_t^{\text{btw}}$). This is because $\delta(t)$ is small and $\beta(t)$ large for long time scales, until the random walk has explored from one cluster to another. That it takes a long time for the random walk to reach global stationarity due to the bottleneck can be argued in terms of graph conductances (Chung, 1997). We sketch a combinatorial argument below, in order to observe how the data dimensionality affects the bottlenecks. For simplicity, we consider just two clusters.

Consider the following model of bottleneck data in $\mathbb{R}^D$. Let $L > 0$ and $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \mathcal{N}$, where $\mathcal{X}_1 = [0,1]^D, \mathcal{X}_2 = [0,1]^D + (L, 0, 0, \ldots, 0)$ are clusters connected at the centers of faces by a cylindrical tube $\mathcal{N}$ of length L and $D-1$ dimensional radius ϵ. Let X be generated by sampling uniformly at random from $\mathcal{X}$. To simplify the analysis, consider a random walk on X such that $x \in X$ transitions to points in $B_\sigma(x)$ with equal probability, and not to any other points, for some $\sigma > 0$.

For $\epsilon < \sigma$, a point x_* at the boundary interface between $\mathcal{X}_1$ and $\mathcal{N}$ is such that most of its near neighbors lie inside the cluster, rather than in the bottleneck. Indeed, up to geometric constants, $\text{vol}_D(X_1 \cap B_\sigma(x_*)) \asymp \sigma^D$, while $\text{vol}_D(\mathcal{N} \cap B_\sigma(x_*)) \asymp$

$\sigma\epsilon^{D-1}$. In particular, the probability of a random walker transitioning from x_* into the bottleneck in one time step is upper bounded by $(\epsilon/\sigma)^{D-1}$, so that the expected number of times a random walk reaches the boundary interface before exiting X_1 is lower bounded by $(\sigma/\epsilon)^{D-1}$. Supposing that the random walker spends time T exploring X_1 before returning to the boundary interface, the expected time it takes for a random walker to exit from X_1 into the bottleneck is $(\sigma/\epsilon)^{D-1}T$. Moreover, once the random walker transitions from X_1 to $\mathcal{N}$, the expected number of steps it takes to cross the bottleneck to get to X_2 may be lower bounded by $\lfloor L/\sigma \rfloor$. Indeed, if all the one-step transitions are optimal in the sense of being distance σ exactly in the direction parallel to $(1, 0, \ldots, 0)$, then at least $\lfloor L/\sigma \rfloor$ steps are required to cross; better estimates on the expected number of steps to cross are also possible (Levin et al., 2009).

In particular, the random walker will, with high probability, exit X_1 then return to X_1 $\Omega(\lfloor L/\sigma \rfloor)$ times, before exiting X_1 and making it all the way to X_2. Thus, the expected time to transition from an arbitrary starting point in X_1 to any point in X_2 may be crudely estimated to be $\Omega(\lfloor L/\sigma \rfloor(\sigma/\epsilon)^{D-1}T)$, which is large when L is large and ϵ is small, i.e. the bottleneck is long and thin. In particular, D_t^{btw} will remain large for $t \lesssim \lfloor L/\sigma \rfloor(\sigma/\epsilon)^{D-1}T$. Noting that T is comparable to the mixing time on X_1 (and thus X_2) alone, this suggests that there are time scales at which the random walker is at equilibrium on the distinct clusters (albeit with small leakage into the bottleneck at each time step) but at which global equilibrium is not reached. At such time scales, diffusion distances within a cluster are small, and between the two clusters are large.

We remark that this property of diffusion distances for bottlenecks also allows LUND to correctly learn the clusters in the data, since the density is constant on this data set. Indeed, in Figure 4, LUND learns four modes in the four distinct clusters, and consequently labels all points correctly. It is interesting to note that depending on the random sample, LUND may learn three modes in distinct clusters and a fourth mode either in the fourth cluster (as in Figure 4 (a)) or at the intersection of the bottlenecks (a possibility not shown in Figure 4). Whether the fourth mode is in a cluster or in the middle of the bottleneck depends on random fluctuations in the empirical density of the data. Regardless of if fourth mode is in the bottleneck region or in a cluster, the consequent LUND labeling of the data is correct. Indeed, if one of the modes is the bottleneck, then the cluster without a mode will have points assigned the label associated to the bottleneck mode, leading all the clusters to be given correct labels.

4.4 Example: A Simple Gaussian Mixture Model

The major parameters controlling the estimates in Theorem 4.12 are δ, λ_{K+1}, and κ. To illustrate the key quantities of this theorem, we consider the simple example of a mixture of Gaussians $\mu_G = \frac{1}{2}\mathcal{N}(\mathbf{x}_1, \Sigma) + \frac{1}{2}\mathcal{N}(\mathbf{x}_2, \Sigma)$ in $\mathbb{R}^2$ with diagonal isotropic covariance matrix $\Sigma = \frac{1}{10}\mathbf{I}$. We construct the diffusion transition matrix $\mathbf{P}$ using the Gaussian kernel with $\sigma = .2$, as described in Section 2.2.

As $\|\mathbf{x}_1 - \mathbf{x}_2\|_2$ increases, Theorem 4.12 becomes more informative. In Figure 7, samples are drawn from μ_G with different amounts of separation (and hence different $\delta, \lambda_{K+1}, \kappa$ values) and we show the dependence on ϵ of the bounds

$$\overline{D_t^{\text{in}}} = 2\frac{\epsilon}{\sqrt{n}}\gamma(t), \underline{D_t^{\text{btw}}} = 2\min_{y\in X}\|\mathbf{s}^\infty(y,\cdot)\|_{\ell^2(\nu)} - 2\frac{\epsilon}{\sqrt{n}}\gamma(t).$$

and the permissible time interval $[\ln\left(\frac{2\kappa}{\epsilon}\right)/\ln\left(\frac{1}{\lambda_{K+1}}\right), \frac{\epsilon}{2\delta}]$. For Theorem 4.12 to be meaningful, ϵ must be such that simultaneously $\overline{D_t^{\text{in}}} < \underline{D_t^{\text{btw}}}$ and $\ln\left(\frac{2\kappa}{\epsilon}\right)/\ln\left(\frac{1}{\lambda_{K+1}}\right) < \frac{\epsilon}{2\delta}$. As $\epsilon \to 0^+$, $\overline{D_t^{\text{in}}} < \underline{D_t^{\text{btw}}}$ holds if the clusters are internally well-connected and separated, as articulated in (4.13), while $\ln\left(\frac{2\kappa}{\epsilon}\right)/\ln\left(\frac{1}{\lambda_{K+1}}\right) \to \infty$ and $\frac{\epsilon}{2\delta} \to 0^+$; a similar but reversed dichotomy occurs as $\epsilon \to \infty$. Figure 7 illustrates this tension between the hypotheses of Theorem 4.12 and the strength of its conclusion.

4.5 Robustness to Geometric Deformations

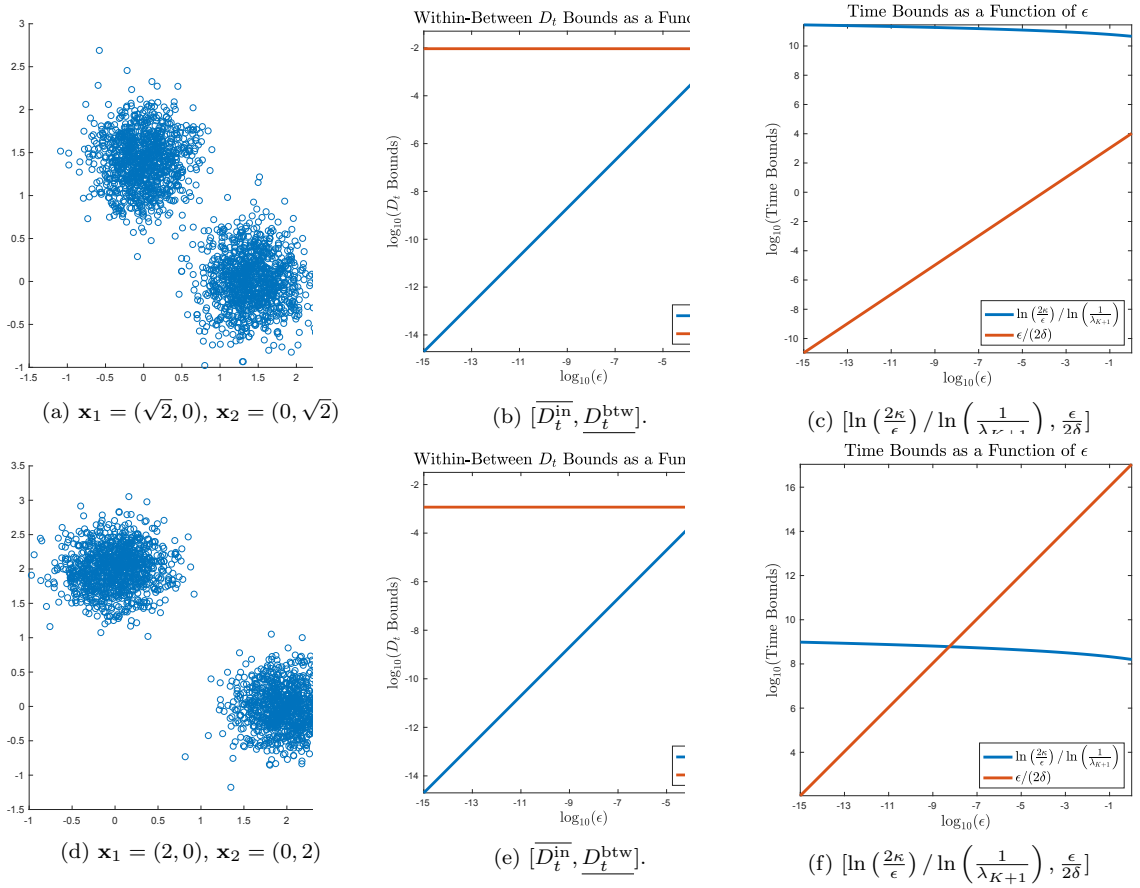
LUND depends mainly on the intrinsic geometric constants δ and λ_{K+1}, so clustering performance with LUND is robust to small amounts of geometric deformation, for example the action of a bi-Lipschitz map. This is illustrated in Figure 8, in which two circular clusters in $\mathbb{R}^2$ are distorted by stretching the second coordinate. The results of LUND as a function of t and the amount of stretching are shown. It can be seen that the more the clusters are stretched, the larger t needs to be to achieve good accuracy. This is because as the clusters are stretched, the parameter λ_{K+1} increases towards 1, meaning that t must be larger to ensure that mesoscopic equilibria on the distinct clusters are reached. Note that the relationship between domain geometry and the eigenvalues of the Laplacian is well-studied in classical settings (Szegö, 1954; Weinberger, 1956), and its role in graph-based clustering is noted in contexts other than LUND (Ng et al., 2002; Arias-Castro, 2011; Schiebinger et al., 2015; Little et al., 2017).

We note that transformations that push the clusters closer together can strongly impact LUND, since this will dramatically change the δ value. This is true of all clustering algorithms; the primary benefit of LUND compared to a range of existing methods is its robustness to distorting the shape of the clusters.

4.6 Relationship Between Time and Scaling Parameter in Diffusion Distances

The Markov chain underlying diffusion distances is typically constructed using the heat kernel $\mathcal{K}(x,y) = \exp(-\|x-y\|_2^2/\sigma^2)$, for some choice of (spatial) scale parameter σ. Once $\mathbf{P} = \mathbf{P}_\sigma$ is constructed, the time parameter t enters. For data sampled from a common manifold, there exists an asymptotic relationship between t and σ as $n \to \infty$:

$$\lim_{\sigma\to 0}\mathbf{P}_\sigma^{T_0/\sigma} = e^{-T_0\Delta}, \tag{4.18}$$



(a) $\mathbf{x}_1 = (\sqrt{2}, 0)$, $\mathbf{x}_2 = (0, \sqrt{2})$

(b) $[\overline{D_t^{\text{in}}}, \underline{D_t^{\text{btw}}}]$.

(c) $[\ln\left(\frac{2\kappa}{\epsilon}\right)/\ln\left(\frac{1}{\lambda_{K+1}}\right), \frac{\epsilon}{2\delta}]$

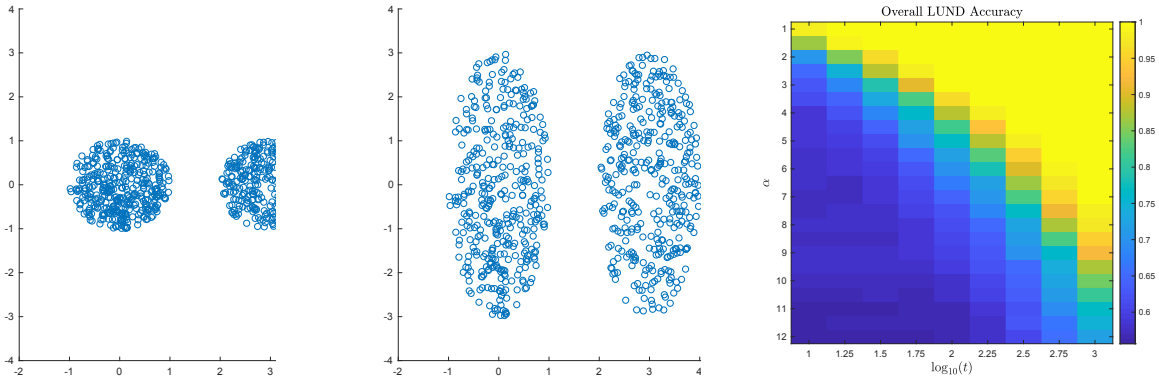
(d) $\mathbf{x}_1 = (2, 0)$, $\mathbf{x}_2 = (0, 2)$

(e) $[\overline{D_t^{\text{in}}}, \underline{D_t^{\text{btw}}}]$.

(f) $[\ln\left(\frac{2\kappa}{\epsilon}\right)/\ln\left(\frac{1}{\lambda_{K+1}}\right), \frac{\epsilon}{2\delta}]$

Figure 7: 2000 data points sampled from μ_G, for two sets of means are shown in (a), (d). In (b), (e), the range between $\overline{D_t^{\text{in}}}$ and $\underline{D_t^{\text{btw}}}$—as a function of ϵ—is shown. Plots (c), (f) show the t interval guaranteed by Theorem 4.12, indicating the existence of a non-empty range of t for which the conclusions of Theorem 4.12 apply whenever the red curve is above the blue curve. As the means $\mathbf{x}_1, \mathbf{x}_2$ move apart, the time interval in which Theorem 4.12 guarantees good separation between the clusters expands. This is makes sense intuitively, since as the clusters move apart, the unsupervised learning problem becomes easier. We remark that the separation for the data in (a) is insufficient for Theorem 4.12 to guarantee a large time range in which $D_t^{\text{in}} < D_t^{\text{btw}}$. This is because the δ constant is large for this data set, since δ is determined by the points in distinct clusters that are nearest, and in particular is a worst-case bound. Relaxing the δ separation condition is discussed in Section 4.3.

where $e^{-T_0\Delta}$ is the infinitesimal generator corresponding to continuous diffusion with canonical time T_0 (Lafon et al., 2006). So, asymptotically as $n \to \infty$, and requiring $\sigma \to 0^+$ and $t = \frac{T_0}{\sigma}$, using (t, σ) is equivalent to using $(Ct, \sigma/C)$ for any constant $C > 0$. This suggests that asymptotically as $n \to \infty$, the performance of LUND with respect to σ, t should be constant if $t\sigma$ is constant as $\sigma \to 0^+$. As we shall see observe in Section 6, working with *finite* data in the *cluster* setting, rather than the asymptotic regime on a common manifold, may lead to more subtle relationships between t and σ.



(a) Circular Data (b) Geometrically Deformed Data (c) LUND Accuracy

Figure 8: In (a), uniform spherical data is shown. In (b), the data in (a) has been geometrically deformed by dilating the second coordinate by a factor of 3. We consider the robustness of LUND to the natural geometric transformation $(x, y) \mapsto (x, \alpha y)$, where α is understood as a stretching parameter. In (c), the results of LUND are shown as a function of t and the dilation factor α. As α increases, the cluster stretches and taking larger t allows LUND to accurately label the data. This is because the random walker requires more time to explore a high elongated cluster than a compact, spherical one.

5. Performance Guarantees for Unsupervised Learning

We consider now how the LUND algorithm (Algorithm 1) performs on data $X = \bigcup_{k=1}^{K} X_k$. Let $p(x)$ be a KDE for $x \in X$, let ρ_t be as in (3.4), and recall $\mathcal{D}_t(x) = p(x)\rho_t(x)$. The LUND algorithm sets the maximizers of $\mathcal{D}_t$ to be the modes of the clusters. Requiring potential modes to have large ρ_t values enforces that modes should be far in diffusion distance from other high-density points, and incorporating $p(x)$ downweighs outliers, which may be far in diffusion distance from their nearest neighbor of higher empirical density.

Theorem 5.1 *Suppose $X = \bigcup_{k=1}^{K} X_k$. If $D_t^{in}/D_t^{btw} < \min(\mathcal{M})/\max(\mathcal{M})$, then the K maximizers $\{\boldsymbol{x}_i^*\}_{i=1}^{K}$ of $\mathcal{D}_t(x)$ are such that $\boldsymbol{x}_i^*$ is a highest empirical density point of X_{k_i} for some permutation $(k_1, \ldots, k_K)$ of $(1, \ldots, K)$.*

Proof We proceed by induction on $1 \leq m \leq K$. Clearly $\mathbf{x}_1^* = \arg\max_{y \in X} p(y)$ is a highest empirical density point of some X_k. Then suppose $\mathbf{x}_1^*, \ldots, \mathbf{x}_m^*, m < K$, have been determined, and are highest empirical density points of distinct classes $X_{k_1}, \ldots, X_{k_m}$. We show that $\mathbf{x}_{m+1}^*$ must be a highest density point among the remaining X_k, $k \notin \{k_1, \ldots, k_m\}$.
First, suppose $x \in X_{k_r}$ for some $r \in \{1, \ldots, m\}$ is any point in the classes already discovered, not of maximal within-class density. Then $\rho_t(x) \leq D_t^{\text{in}}$, since $\mathbf{x}_r^* \in X_{k_r}$ has $p(x) < p(\mathbf{x}_r^*)$, and hence for any $\mathbf{x}$ a highest density point in a cluster not already discovered,

$$\mathcal{D}_t(x) < p(\mathbf{x}_{k_r}^*)\rho_t(x) \leq \max(\mathcal{M}) D_t^{\text{in}} \leq \frac{\max(\mathcal{M})}{\min(\mathcal{M})} \frac{D_t^{\text{in}}}{D_t^{\text{btw}}} p(\mathbf{x})\rho_t(\mathbf{x}) < p(\mathbf{x})\rho_t(\mathbf{x}) = \mathcal{D}_t(\mathbf{x}).$$

Thus, $\mathbf{x}^*_{m+1} \neq x$.
Now, suppose $x \in X_k, k \neq k_1, \dots k_m$. If $x \neq \mathbf{x}$ an empirical density maximizer of X_k, then:

$$\mathcal{D}_t(x) = p(x)\rho_t(x) < p(\mathbf{x})\rho_t(x) \leq p(\mathbf{x})D_t^{\text{in}} < p(\mathbf{x})D_t^{\text{btw}} \leq p(\mathbf{x})\rho_t(\mathbf{x}) = \mathcal{D}_t(\mathbf{x}).$$

Hence, $\mathbf{x}^*_{m+1} \neq x$, and thus $\mathbf{x}^*_{m+1}$ must be among the classwise empirical density maximizers of X_k, $k \notin \{k_1, \dots, k_m\}$. ■

A similar method proves that the ratios of the sorted $\mathcal{D}_t$ values determine the number of clusters K. The problem of estimating the number of clusters is a crucial one, but few methods admit theoretical guarantees; see Little and Byrd (2015) for an overview.

Corollary 5.2 *Let $\{x_{m_i}\}_{i=1}^n$ be the points $\{x_i\}_{i=1}^n$, sorted in non-increasing order: $\mathcal{D}_t(x_{m_1}) \geq \mathcal{D}_t(x_{m_2}) \geq \cdots \geq \mathcal{D}_t(x_{m_n})$. Then:*

(a) $\frac{\mathcal{D}_t(x_{m_j})}{\mathcal{D}_t(x_{m_{j+1}})} \leq \frac{\max(\mathcal{M})}{\min(\mathcal{M})} \frac{\max_{i=1,\cdots,K} \rho_t(x_{m_i})}{\min_{i=1,\cdots,K} \rho_t(x_{m_i})}$ *for $j < K$.*

(b) $\frac{\mathcal{D}_t(x_{m_K})}{\mathcal{D}_t(x_{m_{K+1}})} \geq \frac{\min(\mathcal{M})}{\max(\mathcal{M})} \frac{D_t^{btw}}{D_t^{in}}$.

Proof Statement (a) is immediate from the definition. To see (b), we compute

$$\mathcal{D}_t(x_{m_K}) \geq \min(\mathcal{M})D_t^{\text{btw}} = \frac{\min(\mathcal{M})}{\max(\mathcal{M})} \frac{D_t^{\text{btw}}}{D_t^{\text{in}}} D_t^{\text{in}} \max(\mathcal{M}) \geq \frac{\min(\mathcal{M})}{\max(\mathcal{M})} \frac{D_t^{\text{btw}}}{D_t^{\text{in}}} \mathcal{D}_t(x_{m_{K+1}}).$$

■

Hence if $\frac{D_t^{\text{in}}}{D_t^{\text{btw}}} \frac{\max_{i=1,\cdots,K} \rho_t(x_{m_i})}{\min_{i=1,\cdots,K} \rho_t(x_{m_i})} \ll \left(\frac{\min(\mathcal{M})}{\max(\mathcal{M})}\right)^2$, there will be a sharp drop-off in the values of $\mathcal{D}_t$ after the first K maximizers. Note that the ratio $\min_{i=1,\cdots K} \rho_t(x_{m_i}) / \max_{i=1,\cdots K} \rho_t(x_{m_i})$ will be insignificant unless the clusters are arranged at different scales (i.e. some clusters are very close to each other but far from others). Similarly, $(\min(\mathcal{M})/\max(\mathcal{M}))^2$ will be nearly 1 if the maximal densities of the clusters are comparable. These observations suggest to estimate the number of clusters by selecting a cutoff τ and setting

$$\hat{K} = \inf \left\{ k \,\middle|\, \frac{\mathcal{D}_t(x_{m_k})}{\mathcal{D}_t(x_{m_{k+1}})} > \tau \right\}, \tag{5.3}$$

as in Algorithm 1. Once the modes have been learned correctly, points may be clustered simply by labeling each mode as belonging to its own class, then requiring that every point has the same label as its nearest neighbor in diffusion distance of higher density.

Corollary 5.4 *Suppose $X = \bigcup_{k=1}^{K} X_k$. Let $\{x_{m_i}\}_{i=1}^{n}$ be the points $\{x_i\}_{i=1}^{n}$, sorted so that $\mathcal{D}_t(x_{m_1}) \geq \mathcal{D}_t(x_{m_2}) \geq \cdots \geq \mathcal{D}_t(x_{m_n})$. Then Algorithm 1 labels all points correctly for any τ satisfying*

$$\frac{\max(\mathcal{M})}{\min(\mathcal{M})} \frac{\max_{i=1,\cdots,K} \rho_t(x_{m_i})}{\min_{i=1,\cdots,K} \rho_t(x_{m_i})} < \tau < \frac{\min(\mathcal{M})}{\max(\mathcal{M})} \frac{D_t^{btw}}{D_t^{in}}.$$

Proof By Corollary 5.2, the algorithm correctly estimates $\hat{K}$. Then, by Theorem 5.1, the algorithm correctly learns the empirical density maximizers of each of the $\{X_k\}_{k=1}^{K}$. It remains to show that the subsequent labeling of all points is accurate. For an unlabeled point $x \in X_k$, its nearest diffusion neighbor of higher density, x^*, must be in the same cluster X_k, since $D_t^{\text{in}} < D_t^{\text{btw}}$. Moreover, that point is already labeled as $Y(x^*) = k$, since $p(x^*) \geq p(x)$. Hence, $Y(x) = k$ and by induction, all points are labeled correctly. ■

The dependence on τ is somewhat unsatisfying, and in practice, this quantity can be removed from the inputs of Algorithm 1 by instead setting $\hat{K} = \arg\max_k \mathcal{D}_t(x_{m_k})/\mathcal{D}_t(x_{m_{k+1}})$. This provably detects K accurately by noting that the ratios $\mathcal{D}_t(x_{m_{j+1}})/\mathcal{D}_t(x_{m_j})$ are small for $j > K$ under a range of reasonable assumptions, for example the assumptions that the density of each cluster is bounded away from 0 and the ratio of the minimal and maximal within-cluster diffusion distance is bounded.
Algorithm 2 describes the LUND algorithm in the simpler case that K is known a priori. This algorithm achieves perfect accuracy under milder conditions than Algorithm 1.

Corollary 5.5 *Suppose $X = \bigcup_{k=1}^{K} X_k$ and K is known. If $\frac{D_t^{in}}{D_t^{btw}} < \frac{\min(\mathcal{M})}{\max(\mathcal{M})}$, then Algorithm 2, labels all points correctly.*

Proof This follows from Theorem 5.1, along with $D_t^{\text{in}} < D_t^{\text{btw}}$. ■

6. Numerical Experiments

We return to the motivating data sets of Section 3. The diffusion distances are computed by truncating (2.2) to sum only over the largest (in terms of modulus of the eigenvalues) $M = 100 \ll n$ eigenpairs, and the KDE $p(x)$ uses 100 nearest neighbors with $\sigma_0 = 1$.
We compute a number of statistics on the data to test our theoretical estimates and to verify the efficacy of the proposed algorithm. For the first two data sets we examine, we plot $D_t^{\text{in}}, D_t^{\text{btw}}$ as functions of t, to observe the multitemporal nature of our clustering algorithm. We also compute the theoretical estimates on $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty$ as guaranteed by Theorem 4.8. The tightness of the theoretical estimates is evaluated

by comparing to the empirical values. We moreover plot the diffusion distances from a fixed point for a variety of t values, to illustrate the multitemporal behavior of these distances.

After these evaluations, we cluster the data with the proposed LUND algorithm and compute the accuracy, comparing with spectral clustering and the FSFDPC algorithm. We moreover compute the estimates of K with both the proposed method $\hat{K} = \arg\max_k \mathcal{D}_t(x_{m_k})/\mathcal{D}_t(x_{m_k+1})$ where $\{x_{m_i}\}_{i=1}^n$ are the points $\{x_i\}_{i=1}^n$ sorted so that $\mathcal{D}_t(x_{m_1}) \geq \mathcal{D}_t(x_{m_2}) \geq \cdots \geq \mathcal{D}_t(x_{m_n})$, and spectral clustering eigengap $\hat{K} = \arg\max_i \lambda_{i+1} - \lambda_i$, as a function of the crucial parameters of the respective algorithms. In particular, we evaluate the robustness of spectral clustering methods with respect to σ, and LUND with respect to σ and t. For spectral clustering, we consider the variant in which just the second eigenvector ψ_2 is used (Shi and Malik, 2000), as well as the variant in which the first K eigenvectors $\{\psi_i\}_{i=1}^K$ are used (Ng et al., 2002). All experiments are conducted on randomly generated data, with results averaged over 100 trials.

6.1 Bottleneck Data

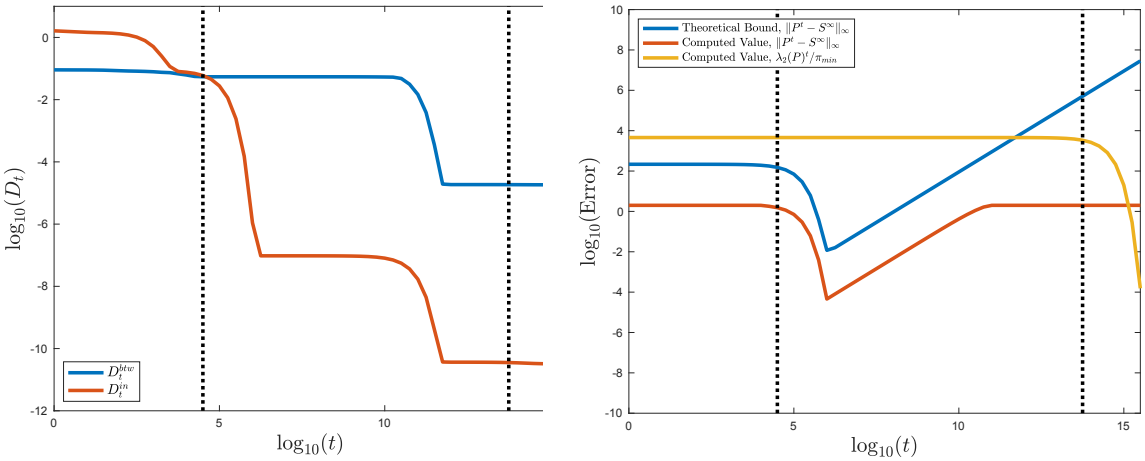
We first analyze the linear, multimodal data set of Figure 2, in which two of the clusters feature two high-density regions, connected by a lower density bottleneck region. Theorem 4.8 upper bounds $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty < \epsilon$ in terms of $\delta, \lambda_{K+1}, \kappa$, which for this data have values $\delta = 6.2697 \times 10^{-8},\ \lambda_{K+1} = 1 - 1.7563 \times 10^{-4},\ \kappa = 2.6738 \times 10^2$ when $\mathbf{P}$ is constructed with $\sigma = .15$. As shown in Figure 9, the theoretical estimate correctly illustrates the overall behavior of the transition from initial distribution, to mesoscopic equilibria, then to a global equilibrium.

The distance from a high-density point across time scales appears in Figure 10. For small time values, the diffusion distance scales similarly to Euclidean distance. However, by time $t = 10^8$, a mesoscopic equilibrium has been reached, and all points in the cluster are rather close together. By $t = 10^{16}$, a global equilibrium has been reached.

6.1.1 Bottleneck Data Clustering Evaluation

Comparisons with spectral clustering appear in Figures 11 and 12. In Figure 12 (a), it is shown that for all values of the spatial scale parameter σ, the eigengap estimated number of clusters $\hat{K}$ is 1, i.e. always incorrect. On the other hand, Figure 12 (b) shows that there is a range of (σ, t) values—mesoscopic in t—for which LUND achieves perfect accuracy. Indeed, after an initial phase in which the number of clusters is estimated as 1, the LUND estimate for K is decreasing in t, corresponding to the mixing of different clusters over time.

The LUND algorithm and FSFDPC are compared in Figure 13. Due to the non-spherical shapes of the clusters, FSFDPC is unable to learn the modes of the data correctly, and consequently assigns modes to the same cluster: the modes learned

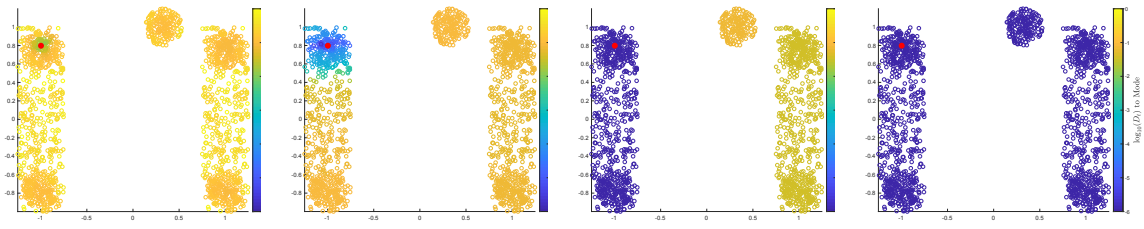


(a) Plot of $D_t^{\text{btw}}, D_t^{\text{in}}$ against t.

(b) Quantities related to mesoscopic and global equilibria.

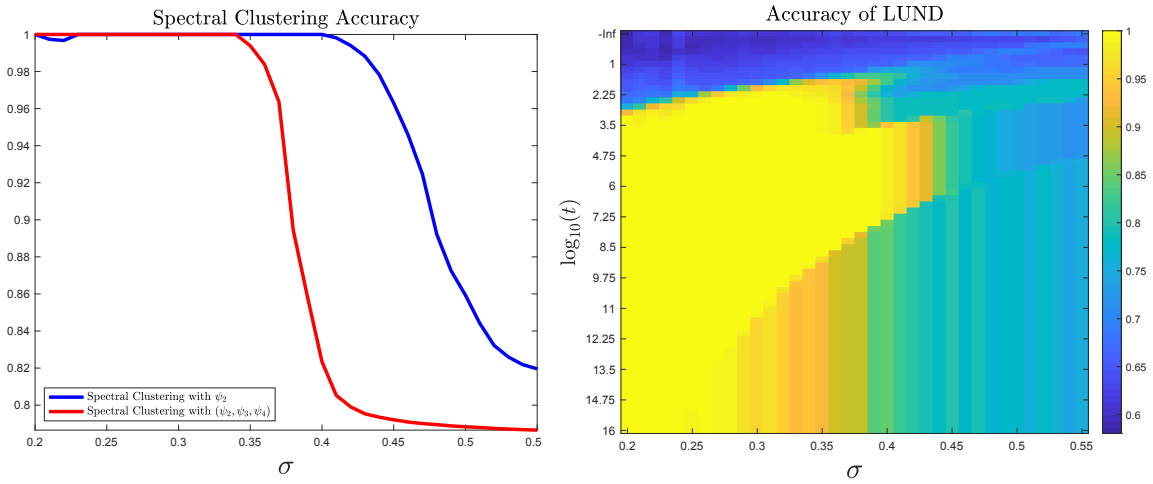
Figure 9: In (a), $D_t^{\text{btw}}, D_t^{\text{in}}$ are plotted against t. For $t < 10^{4.5}$, $D_t^{\text{in}} > D_t^{\text{btw}}$, since for small time, D_t is essentially the same as Euclidean distance. Around $t = 10^{4.5}$, there is a transition, in which $D_t^{\text{in}} \ll D_t^{\text{btw}}$. This corresponds to the Markov chain reaching mesoscopic equilibria in which the chain is well-mixed on each cluster, but not well-mixed globally. The approximate times of convergence towards the mesoscopic and global equilibrium are denoted with dotted black vertical lines. In (b), we plot three quantities against t: the theoretical bound on $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty$ guaranteed by Theorem 4.8; the empirical quantity $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty$; and the empirically computed quantity $\lambda_2(\mathbf{P})^t/\boldsymbol{\pi}_{\min}$, which estimates the distance to the stationary distribution. Notice that $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty$ gets small, both the theoretical bound and the empirical value, around $t = 10^{4.5}$. It then increases. Around $t = 10^{13.75}$, $\lambda_2(\mathbf{P})^t/\boldsymbol{\pi}_{\min}$ decays exponentially to 0, indicating that the global equilibrium has been reached. Note that the theoretical estimate on $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty$ is not tight, though it accurately captures the overall behavior of the quantity with respect to t. Moreover, the global decay estimate $\lambda_2(\mathbf{P})^t/\boldsymbol{\pi}_{\min}$ is somewhat conservative in estimating the mixing time of towards the equilibrium distribution, as can be seen by comparing the second vertical lines in (a) and (b). In order to keep the plots at the same scale, the plots in (a) are extended to be constant for roughly $t > 12$.

by FSFDPC and subsequent labels appear in subfigures (a), (b), respectively. In contrast, LUND learns one mode from each cluster, as shown in (c). Consequently, all points are labeled correctly, as shown in (d).



(a) $t = 0$ (b) $t = 10^4$ (c) $t = 10^8$ (d) $t = 10^{16}$

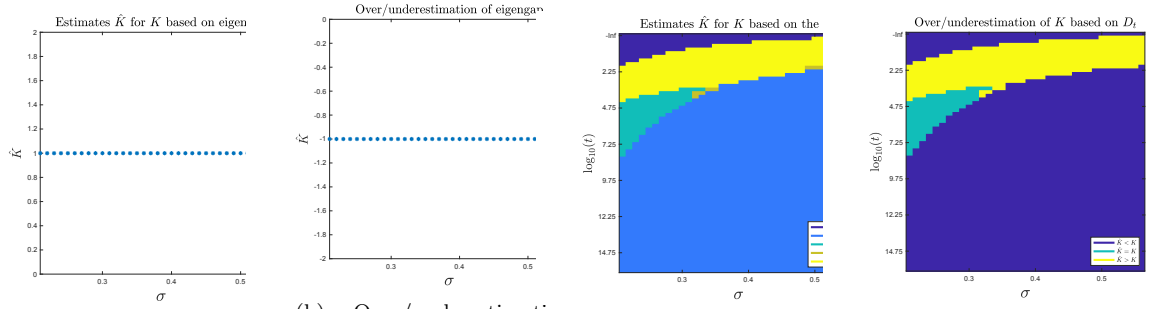
Figure 10: A high-density point is shown in red, and all other points are colored by the D_t distance from this point in $\log_{10}$ scale. The transition from initial distribution (a), to mesoscopic equilibrium (c), to global equilibrium (d), is illustrated as t grows.



(a) Accuracy spectral clustering

(b) Accuracy of LUND

Figure 11: Accuracy of two variations of spectral clustering compared to LUND as functions of σ. While spectral clustering with ψ_2 performs nearly perfectly for $\sigma < .4$, its performance degrades as σ increases. Classical spectral clustering using ψ_2, ψ_2, ψ_4 achieves perfect clustering of the data for roughly $\sigma < .35$. LUND is able to achieve perfect clustering accuracy for a wide range of (σ, t) pairs, mainly for those σ values which allows spectral clustering with just ψ_2 to succeed. As σ increases, the mesoscopic regime in which perfect accuracy is achieved shrinks before disappearing entirely around $\sigma = .45$. In this data, spectral clustering with just ψ_2 performs about as well as LUND in terms of accuracy, assuming K is known.



(a) Estimates of $\hat{K}$ using eigengap statistic.

(b) Over/underestimation of $\hat{K}$ using eigengap statistic.

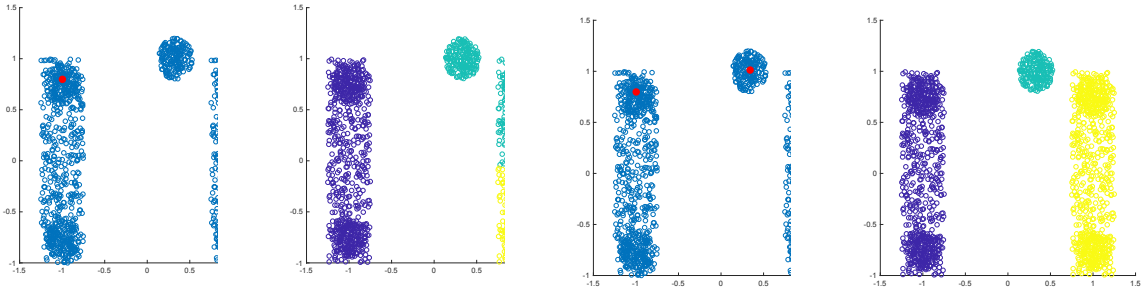
(c) Estimates of $\hat{K}$ using $\mathcal{D}_t$ statistic.

(d) Over/underestimation of $\hat{K}$ using $\mathcal{D}_t$ statistic.

Figure 12: In (a), we see the estimates of $\hat{K}$ using the eigengap statistic, as a function of spatial scale parameter σ. The eigengap consistently estimates $\hat{K} = 1 < 3 = K$, indicating that the multimodal nature of this data is too complicated for the spectral clustering eigengap to handle. A quantized version of these estimates is shown in (b), in which entry 0 indicates correct estimation, -1 indicates $\hat{K} < K$ and 1 indicates $\hat{K} > K$. There is a regime of (σ, t) values in which LUND correctly estimates the number of clusters, as shown in (c) and (d). After an initial time, this regime is essentially monotonic decreasing in t, and the mesoscopic region in which $\hat{K} = K$ is decreasing in σ.

6.2 Nonlinear Data

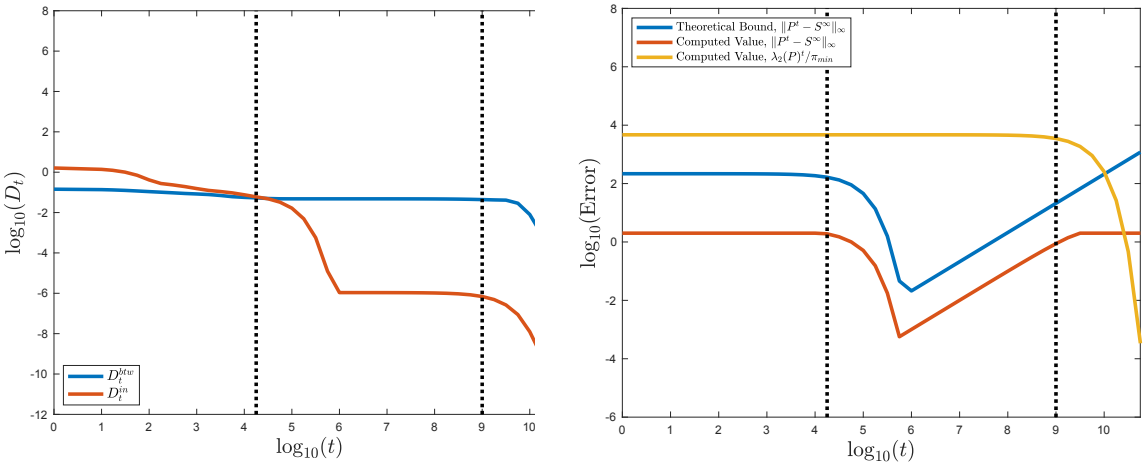
We now consider the nonlinear multimodal data of Figure 2 (b). The innermost circle is filled-in, and has only one high-density region. It is surrounded by two circles, each with two high-density regions connected by low-density regions. The paths connecting



(a) Learned modes, FSFDPC (b) Learned FSFDPC (c) Learned modes, LUND (d) Learned labels, LUND

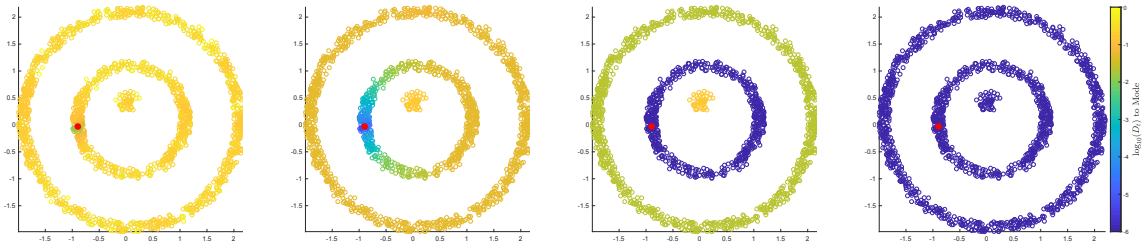
Figure 13: Comparison of FSFDPC to LUND. In (a), the modes learned from FSFDPC—with Euclidean distances—are plotted. Due to the eloganted, non-spherical nature of the data, the modes are learned incorrectly. The subsequent labels, shown in (b), illustrate FSFDPC is not able to capture the structure of this data. In (c), the modes learned from LUND are shown. One mode is learned from each cluster, which allows for a correct labeling of all data points with LUND, as shown in (d). LUND used parameters $(\sigma, t) = (.15, 10^6)$ for these data.

antipodal points on the outer circles are long, which suggests these sets will have low conductance. In the context of Theorem 4.8, the parameters for this data have values $\delta = 1.7225 \times 10^{-4}, 1 - \lambda_{K+1} = 6.8350 \times 10^{-5}, \kappa = 2.655 \times 10^2$ with $\sigma = .2$. Comparison of theoretical and empirical estimates appear in Figure 14, and the diffusion distances from one of the high-density points appear in Figure 15, illustrating the transition



(a) Plot of $D_t^{\text{btw}}, D_t^{\text{in}}$ against t. (b) Quantities related to mesoscopic and global equilibria.

Figure 14: In (a), we plot $D_t^{\text{btw}}, D_t^{\text{in}}$ against t. For roughly $t < 10^{4.25}$, $D_t^{\text{in}} > D_t^{\text{btw}}$; around $t = 10^{4.25}$, there is a transition, in which $D_t^{\text{in}} \ll D_t^{\text{btw}}$. This corresponds to the Markov chain reaching mesoscopic equilibria in which the chain is well-mixed on each cluster, but not well-mixed globally. The approximate times of convergence towards the mesoscopic and global equilibrium are denoted with dotted black vertical lines. In (b), we plot three quantities against t: the theoretical bound on $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty$ guaranteed by Theorem 4.8; the empirically computed quantity $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty$; and the empirically computed quantity $\lambda_2(\mathbf{P})^t/\boldsymbol{\pi}_{\min}$, which estimates the distance to the stationary distribution. Notice that $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty$ gets small, both the theoretical bound and the empirical value, around $t = 10^{4.25}$. It then increases. Around $t = 10^9$, $\lambda_2(\mathbf{P})^t/\boldsymbol{\pi}_{\min}$ decays exponentially to 0, indicating that the global equilibrium has been reached.

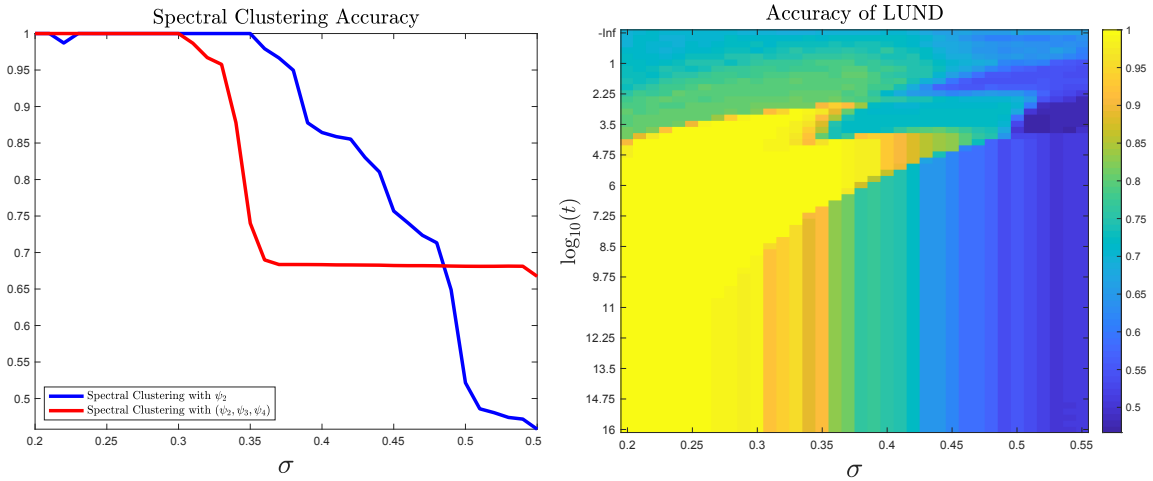


(a) $t = 0$ (b) $t = 10^4$ (c) $t = 10^8$ (d) $t = 10^{16}$

Figure 15: A high-density point is shown in red, and all other points are colored by D_t distance from this point in $\log_{10}$ scale. The transition from initial distribution (a), to mesoscopic equilibrium (c), to global equilibrium (d), is illustrated as t grows.

6.2.1 Nonlinear Data Clustering Evaluation

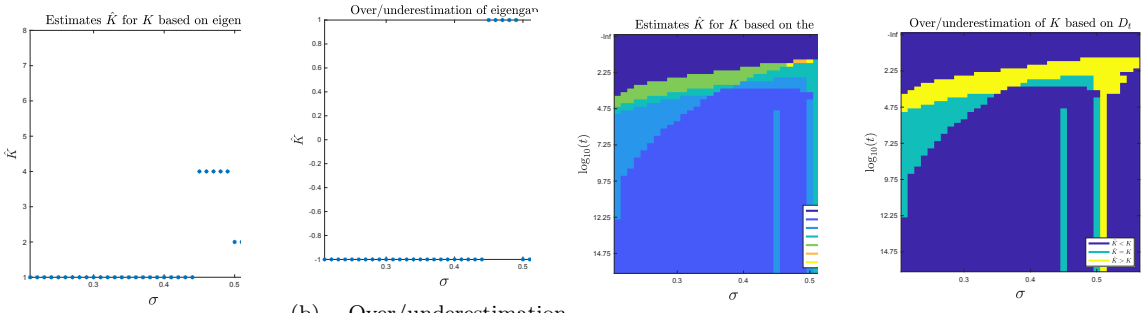
LUND is compared with the two spectral clustering variants in Figures 16 and 17. In terms of overall accuracy, LUND with the correct choice of t outperforms both methods of spectral clustering—using ψ_2 only and using ψ_2, ψ_3, ψ_4—for a range of σ values. The strong performance of LUND in the mesoscopic range, away from $t = 0, t = \infty$, confirms the theoretical results, and demonstrates LUND's flexibility compared to classical spectral methods. Beyond accuracy, the LUND estimator for K is empirically effective for a range of (σ, t) values, while the eigengap is much less effective.



(a) Accuracy spectral clustering (b) Accuracy of LUND

Figure 16: Accuracy of two variations of spectral clustering compared to LUND as functions of σ. While spectral clustering with ψ_2 performs well for very small σ, its performance degrades as σ increases; classical spectral clustering using ψ_2, ψ_3, ψ_4 performs similarly though for a smaller range of σ. LUND is able to achieve perfect clustering accuracy for a wide range of (σ, t) pairs, in particular for pairs (σ, t) such that spectral clustering fails. As σ increases, the mesoscopic regime in which perfect accuracy is achieved shrinks before disappearing entirely around $\sigma = .4$. LUND outperforms spectral clustering with (ψ_2, ψ_3, ψ_4) roughly for $\sigma \in (.3, .4)$, and outperforms spectral clustering with ψ_2 alone roughly for $\sigma \in (.35, .4)$.

In Figure 18, LUND is compared to FSFDPC. LUND correctly learns the modes of the data and labels points correctly, as shown in (c), (d). FSFDPC, however, fails to



(a) Estimates of $\hat{K}$ using eigengap statistic.

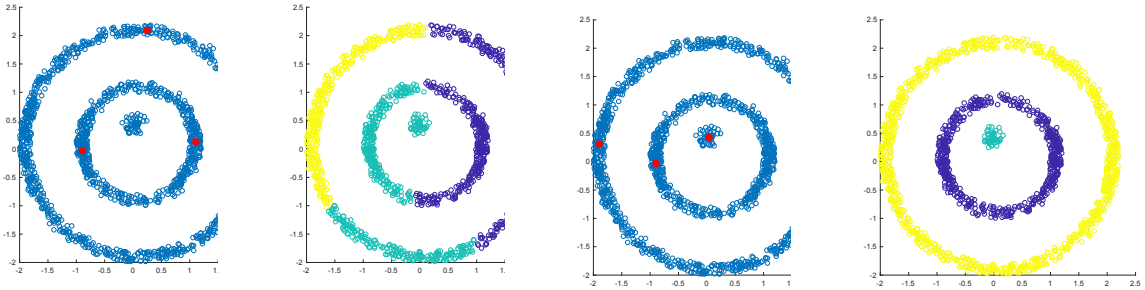
(b) Over/underestimation of $\hat{K}$ using eigengap statistic.

(c) Estimates of $\hat{K}$ using $\mathcal{D}_t$ statistic.

(d) Over/underestimation of $\hat{K}$ using $\mathcal{D}_t$ statistic.

Figure 17: In (a) and (b), we see the eigengap consistently misestimates the number of clusters in the data, first estimating $\hat{K} = 1$, before oscillating between various numbers of clusters, including ending on $\hat{K} = 3$ for one value of σ. LUND is able to achieve an accurate estimate for $\hat{K}$ for a range of (σ, t) values, with generally more t values yielding a correct estimate for smaller σ. We note that the vertical lines in (c), (d) around $\sigma = .45$ and $\sigma = .5$ that deviate from the surrounding values are due to using the mode as the summary statistic across the trials. If the mean is used, a much smoother plot emerges, though one that is less meaningful in terms of estimating the number of clusters.

learn the modes of the data correctly, leading to erroneous labeling—see subfigures (a), (b) respectively.



(a) Learned modes, FSFDPC

(b) Learned FSFDPC

(c) Learned modes, LUND

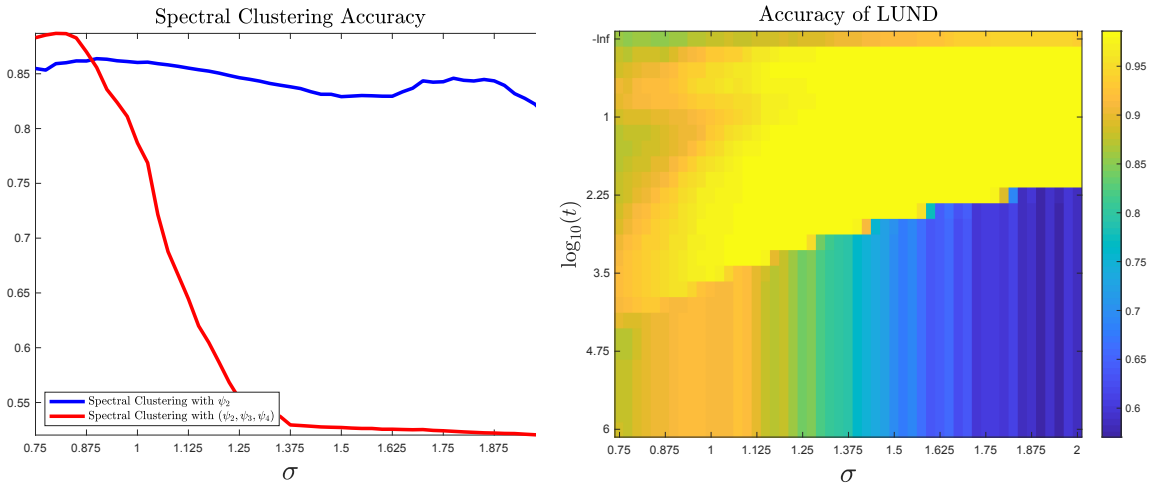
(d) Learned labels, LUND

Figure 18: In (a), the modes learned from FSFDPC—with Euclidean distances—are plotted. The nonlinear nature of the data causes the modes to be learned incorrectly. The subsequent labels, shown in (b), illustrate FSFDPC is not able to capture the structure of this data. In (c), the modes learned from LUND are shown. One mode is learned from each cluster, which allows for a correct labeling of all data points with LUND, as shown in (d). LUND used parameters $(\sigma, t) = (.175, 10^5)$ for these data.

6.3 Gaussian Data

As a final synthetic example, we consider the Gaussians of Figure 1, which were constructed by Nadler and Galun (2007) to be data on which both variants of spectral clustering fail. These data are not sufficiently separated for Theorem 4.12 to apply, but LUND still is able to perform well, owing to the incorporation of density, which allows to easily estimate the modes of the data. Comparisons to spectral clustering in terms of overall accuracy are reported in Figure 19. It is also enlightening to consider

performances of LUND and spectral clustering in terms of *average accuracy*, in which the overall accuracy on each of the clusters is computed separately, and these class-wise accuracies are then averaged. Compared to the overall accuracy measure, the average accuracy measure discounts large clusters and increases the significance of small clusters. Results for average accuracy are in Figure 20



(a) Accuracy of spectral clustering

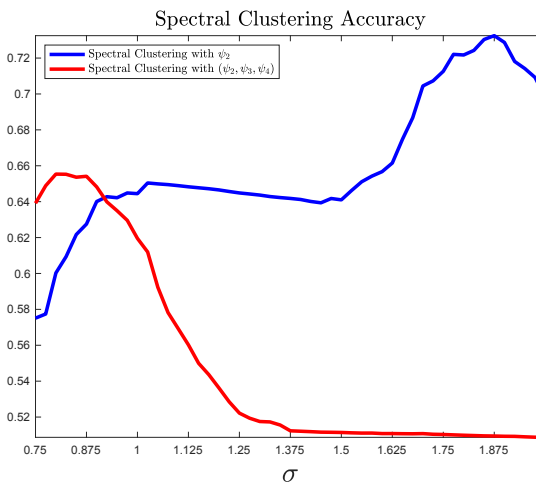
(b) Accuracy of LUND

Figure 19: The *overall accuracy* of the two spectral clustering variants, as well as LUND, are shown for the Gaussian data. In terms of overall accuracy, LUND is able to achieve near-perfect results for a range of parameter values. Nearly all errors made were due to a point being generated from one Gaussian and landing very close to another Gaussian, which is essentially an unavoidable identifiability issue from which any unsupervised clustering method would suffer. Neither of the spectral clustering methods is able to match LUND's performance, which can be attributed to fundamental issues with the use of only the first small number of eigenvectors when performing spectral clustering, as shown by Nadler and Galun (2007) and illustrated in Figure 1.

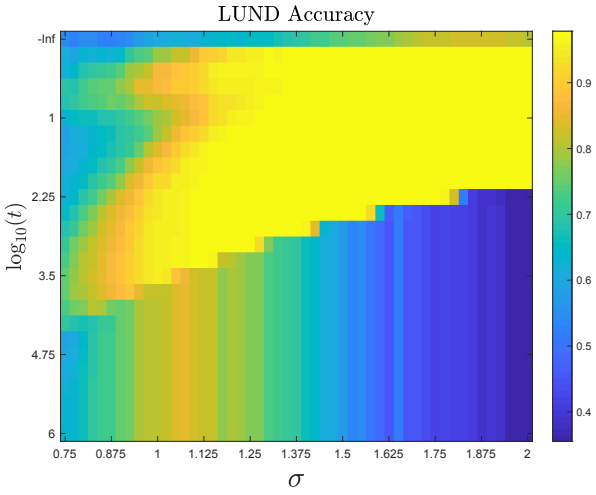
In Figure 21, LUND is compared to FSFDPC. LUND correctly learns the modes of the data and labels points correctly, as shown in (c), (d). FSFDPC also learns the modes correctly, due to the unimodality of the Gaussian clusters and their isotropic covariance matrices.

6.4 Experimental Conclusions

In all three synthetic examples, LUND performs well. On the bottleneck data, it gives the same accuracy as spectral clustering with ψ_2 but better estimates on $\hat{K}$; on the nonlinear data, it gives the best range of accuracies with respect to σ, while also giving the best estimates of $\hat{K}$; on the Gaussian data, LUND performs as well as FSFDPC while both spectral methods fail. These results suggest that LUND combines the best properties of spectral clustering with density-based clustering, while enjoying theoretical guarantees. We remark that extensive experiments with a variant of LUND adapted to high-dimensional images were performed on real hyperspectral image data (Murphy and Maggioni, 2018a,b, 2019b), demonstrating the competitiveness of LUND with a range of benchmark and state-of-the-art clustering algorithms.

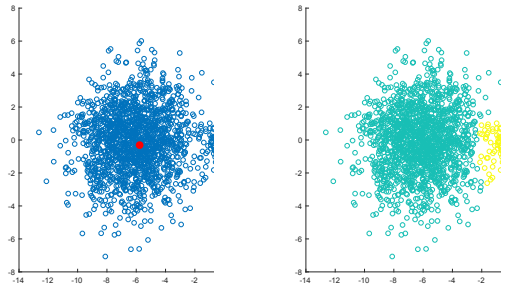


(a) Accuracy of spectral clustering



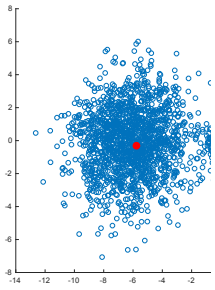
(b) Accuracy of LUND

Figure 20: The *average accuracy* of the two spectral clustering variants, as well as LUND, are shown for the Gaussian data. The results are qualitatively similar to overall accuracy, but with reduced performance for spectral clustering, since most of the errors made by the spectral clustering variants are on the small cluster, which is washed out by spectral clustering. LUND achieves essentially perfect performance for a range of parameter values, excepting identifiability issues.

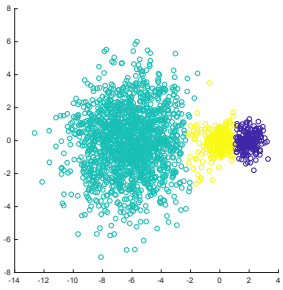


(a) Learned modes, FSFDPC

(b) Learned FSFDPC



(c) Learned modes, LUND



(d) Learned labels, LUND

Figure 21: For the Gaussian data, both LUND and FSFDPC learn the data modes correctly, and are subsequently able to label the data with high accuracy. The lack of difference in their comparative performances is attributed to the fact that the data in this case are Gaussians with isotropic covariance matrices, and in particular have simple spherical supports, which confers diffusion distances little advantage compared to Euclidean distances. LUND used parameters $(\sigma, t) = (1, 10^3)$ for these data.

As shown in Figures 11, 12, 16, 17, 19, 20, the relationship between σ and t is not as simple as suggested by equation (4.18). Indeed, in the non-asymptotic case, and in particular in the case of well-separated clusters, the relationship between t and σ does not obey a strict exponential relationship, as suggested by (4.18). Instead, t appears to interact with scales *locally* on each cluster, as can be seen by the bifurcations in these plots. Gaining a complete understanding of the relationship between σ and t in the cluster case is a topic of ongoing research.

6.5 Computational Complexity

The proposed algorithm enjoys essentially linear computational complexity. This is achieved through the indexing structure *cover trees* (Beygelzimer et al., 2006), which allows for efficient nearest neighbor searches under the assumption that data has low-dimensional structure. Indeed, for data $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$, the computation of each of the n data point's k_{nn} nearest neighbors can be achieved at a cost of $O(k_{\text{nn}}DC^d n\log(n))$, where d is the intrinsic dimension of the data. This allows for the computation of each point's empirical density estimate $p(x)$ in $O(k_{\text{nn}}DC^d n\log(n))$, where k_{nn} is the number of nearest neighbors used in the density estimate. In addition, if the Markov chain is computed not on a fully connected graph, but on a k_{nn} nearest neighbors graph, the cost of computing diffusion distances with the eigenvector approximation is $O(k_{\text{nn}}DC^d n\log(n) + k_{\text{nn}}M^2 n)$, where M eigenvectors are used in the approximation. In the case that $k_{\text{nn}} = O(\log(n)), M = O(1)$, these complexities simplify to $O(DC^d n\log(n)^2)$. Computing all $\rho_t(x)$ values is $O(n\log(n))$, under the assumption that, except for the $O(\log(n))$ class modes, each point has among its $O(\log(n))$ nearest neighbors a point of higher empirical density. Subsequent sorting of p and D_t are $O(n\log(n))$, so the overall algorithm is $O(DC^d n\log^2(n))$, which is linear in n, up to logarithmic factors.

6.6 Discussion and Tuning of Parameters

The LUND algorithm depends crucially on the parameters σ, t, and also potentially on the cutoff parameter τ used to estimate K as in (5.3). Our experiments have demonstrated that LUND is robust to σ and t, and that the parameter τ may be dispensed with by estimating $\hat{K} = \arg\max_k \mathcal{D}_t(x_{m_k})/\mathcal{D}_t(x_{m_{k+1}})$.

We remark that σ may be set *locally* in an automated way by replacing a global σ with a point-dependent σ based on distances to near neighbors (Zelnik-Manor and Perona, 2005). This approach modifies the construction of $\mathbf{P}$ by setting $\mathbf{W}_{ij} = \exp(-\|x_i - x_j\|_2^2/\sigma_i\sigma_j)$, where $\sigma_i = \|x_i - NN(x_i, k_{\text{NN}})\|_2$, where $NN(x_i, k_{\text{NN}})$ is the k_{NN}^{th} Euclidean nearest neighbor of x_i. Which nearest neighbor to use, k_{NN}, is a function of the data dimension, but the approach is practically quite robust to this parameter. This has the effect of adapting to the local density near each point. Moreover, it suggests that if the underlying density is smooth and bounded away from 0, then for all x_i, there is a natural scaling $\sigma_i \to 0^+$ as $n \to \infty$, with rate depending on the data dimensionality.

Regarding t, this parameter determines the time scale at which the diffusion is run. The theoretical results guarantee that if a certain clustering of the data admits suitable separation and geometric properties, then there will be choices of t for which LUND accurately labels the data. In practice, choosing t without supervision can be performed in several ways. One approach is to choose the t which maximizes the

ratio between successive values of $\mathcal{D}_t(x)$:

$$t^* = \arg\max_t \left\{ \arg\max_k \mathcal{D}_t(x_k)/\mathcal{D}_t(x_{k+1}) \right\}.$$

This is analogous to the method for selecting the optimal parameters for ultrametric spectral clustering in Little et al. (2017). A different approach is to select t^* such that clustering with LUND at time t^* is "stable" (Wan and Meila, 2016; Meila, 2019) with respect to small perturbations in the underlying data. Relatedly, it is natural to select t such that the clustering results are stable to small perturbations in t.
We remark that in some sense, multiple choices of t, leading to different clusterings, may be reasonable. This is because there may be hierarchical cluster structure in the data, parametrized by t. Developing extensions and theoretical analyses of LUND in this context is the subject of ongoing research.

7. Conclusions and Future Work

In this article, new methods for bounding diffusion distances, based on nearly-reducible Markov chains, are deployed to provide sufficient conditions under which clustering of data can be guaranteed. The theoretical results rigorously show that diffusion distances exhibit multitemporal behavior, even in the case that clusters have multiple regions of high-density or nonlinear support. These estimates on diffusion distance allow to prove performance guarantees on the LUND algorithm. This may be interpreted as a critique of the popular FSFDPC algorithm, for which theoretical guarantees require unrealistic assumptions on the shapes of the clusters. Numerical experiments on bottleneck, nonlinear, and Gaussian data indicate that the theoretical results correspond with empirical performance, and that LUND enjoys advantages of both spectral clustering and FSFDPC while tempering their weaknesses.
While the results presented in this paper for diffusion distances are, we believe, novel and useful for developing performance guarantees for unsupervised learning, they fall short of a full finite sample analysis. It is of interest to understand how the estimates on δ and λ_{K+1}, which govern the estimates of Theorem 4.12, and consequently the performance guarantees for clustering, scale with the number of sample points. Developing such precise estimates would require new mathematical methods for analyzing the spectra of random operators on graphs. Such an analysis is suggested by recent works in discrete-to-continuum spectral analysis (Garcia Trillos et al., 2018), though handling the factor $(\mathbf{I} - \mathbf{P}_{ii})^{-1}$ may provide for new challenges.
As remarked in Section 6, diffusion distances are Euclidean distances in a new coordinate basis, given by the (right) eigenvectors of $\mathbf{P}$. A different approach to proving the clustering properties of D_t with respect to time would be to show that different eigenvectors localize on particular clusters, and show that there are gaps in the eigenvalues λ_ℓ which account for the emergence of mesoscopic equilibria. This approach is related to the analysis of eigenvectors corresponding to small eigenvalues for the symmetric Laplacian (Shi et al., 2009), and may provide new insights.

In Section 4.6, the asymptotic relationship between σ and t for diffusion distances is discussed (Lafon et al., 2006). As shown in Section 6, LUND exhibits a more delicate relationship between σ, t in the finite sample case. This is potentially due to the incorporation of empirical density into the diffusion geometry framework that LUND uses to estimate the data modes. Analyzing this phenomenon mathematically is a topic of future research.

In LUND, the modes are learned sequentially, by selecting the K maximizers of $\mathcal{D}_t$. It is of interest to consider whether jointly selecting the modes, based on a criterion that optimizes over K points simultaneously, leads to different and in some cases improved clustering results in particular situations. Such a formulation may force the modes to be better spread throughout the data, and in particular would force the modes in the bottleneck data in Figure 4 to localize on the clusters, regardless of the fluctuations in empirical density. For this data, it does not affect the clustering accuracy if the fourth mode is in the middle of the bottleneck or in a cluster, but it is intuitive that the modes should localize on the four cluster cores rather than anywhere along the bottleneck. Developing a joint mode selection criterion may enforce this.

The proposed method also lends itself to the semisupervised setting of *active learning* (Chapelle et al., 2006; Dasgupta, 2011), in which the user is allowed to query a small number of points for labels. By estimating which points are most likely to be modes of clusters, the LUND algorithm presents natural candidates to query for labels. Recent work has suggested this approach is feasible (Murphy and Maggioni, 2018b, 2019a; Maggioni and Murphy, 2019), and it is an ongoing research effort to optimally integrate diffusion geometry into an active sampling scheme.

Acknowledgments

We thank the three anonymous reviewers, whose comments and insights greatly improved the manuscript. This project was partially funded by NSF-DMS-125012, NSF-DMS-1724979, NSF-DMS-1708602, NSF-ATD-1737984, AFOSR FA9550-17-1-0280, NSF-IIS-1546392, NSF-DMS 1912737, and NSF-DMS 1924513.

Appendix A. Proof of Theorem 4.8

Notice $\|\mathbf{P}^t - \mathbf{S}^\infty\|_\infty \leq \|\mathbf{P}^t - \mathbf{S}^t\|_\infty + \|\mathbf{S}^t - \mathbf{S}^\infty\|_\infty$. For all $t \geq 0$, $\mathbf{P}^t - \mathbf{S}^t = \sum_{i=1}^t \mathbf{S}^{t-i}(\mathbf{P} - \mathbf{S})\mathbf{P}^{i-1}$, so that

$$\|\mathbf{P}^t - \mathbf{S}^t\|_\infty = \left\| \sum_{i=1}^t \mathbf{S}^{t-i}(\mathbf{P} - \mathbf{S})\mathbf{P}^{i-1} \right\|_\infty \leq \sum_{i=1}^t \|\mathbf{S}^{t-i}\|_\infty \|(\mathbf{P} - \mathbf{S})\|_\infty \|\mathbf{P}^{i-1}\|_\infty = t\|(\mathbf{P} - \mathbf{S})\|_\infty \leq t\delta.$$

To bound $\|\mathbf{S}^t - \mathbf{S}^\infty\|_\infty$, notice that after diagonalizing $\mathbf{S}$,

$$\mathbf{S}^t = \mathbf{Z} \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{D}^t \end{bmatrix} \mathbf{Z}^{-1}, \quad \mathbf{S}^\infty = \mathbf{Z} \begin{bmatrix} \mathbf{I}_K & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Z}^{-1},$$

where $\mathbf{D}$ is a diagonal matrix with $\lambda_{K+1}, \lambda_{K+2}, \ldots, \lambda_n$ on the diagonal. Hence, $\|\mathbf{S}^t - \mathbf{S}^\infty\|_\infty \leq \|\mathbf{Z}\|_\infty \lambda_{K+1}^t \|\mathbf{Z}^{-1}\|_\infty = \kappa \lambda_{K+1}^t$, as desired. The second result of the theorem follows similarly.

References

D. Aldous and J. Fill. Reversible Markov chains and random walks on graphs, 2002.

R. Andersen and Y. Peres. Finding sparse cuts locally using evolving sets. In *Symposium on Theory of Computing (STOC)*, pages 235–244. ACM, 2009.

R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *Symposium on Foundations of Computer Science (FOCS)*, pages 475–486. IEEE, 2006.

R. Andersen, F. Chung, and K. Lang. Local partitioning for directed graphs using pagerank. *Internet Mathematics*, 5(1-2):3–22, 2008.

E. Arias-Castro. Clustering based on pairwise distances when the data is of mixed dimensions. *IEEE Transactions on Information Theory*, 57(3):1692–1706, 2011.

E. Arias-Castro, G. Chen, and G. Lerman. Spectral clustering based on local linear approximations. *Electronic Journal of Statistics*, 5:1537–1587, 2011.

E. Arias-Castro, G. Lerman, and T. Zhang. Spectral clustering based on local PCA. *Journal of Machine Learning Research*, 18(9):1–57, 2017.

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Symposium on Discrete Algorithms (SODA)*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

A. Athreya, D.E. Fishkind, M. Tang, C.E. Priebe, Y. Park, J.T. Vogelstein, K. Levin, V. Lyzinski, and Y. Qin. Statistical inference on random dot product graphs: a survey. *The Journal of Machine Learning Research*, 18(1):8393–8484, 2017.

M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

M. Belkin and P. Niyogi. Towards a theoretical foundation for Laplacian-based manifold methods. In *International Conference on Computational Learning Theory (COLT)*, pages 486–500. Springer, 2005.

M. Belkin and P. Niyogi. Convergence of Laplacian eigenmaps. In *Advances in Neural Information Processing Systems (NIPS)*, pages 129–136, 2007.

A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. In *International Conference on Machine Learning (ICML)*, pages 97–104. ACM, 2006.

S. Botelho-Andrade, P.G. Casazza, D. Cheng, and R. Tran. The exact constant for the ℓ_1—ℓ_2 norm inequality. *Mathematical Inequalities & Applications*, 22(1):59–64, 2019.

J.E. Chacón. Clusters and water flows: a novel approach to modal clustering through morse theory. *arXiv preprint arXiv:1212.1384*, 2012.

O. Chapelle, B. Scholkopf, and A. Zien. *Semi-supervised Learning*. MIT Press, 2006.

G. Chen and G. Lerman. Foundations of a multi-way spectral clustering framework for hybrid linear modeling. *Foundations of Computational Mathematics*, 9(5):517–558, 2009a.

G. Chen and G. Lerman. Spectral curvature clustering (SCC). *International Journal of Computer Vision*, 81(3):317–330, 2009b.

F. Chung. *Spectral Graph Theory*, volume 92. American Mathematical Soc., 1997.

R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

R.R Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences of the United States of America*, 102(21):7426–7431, 2005.

R.R. Coifman, I.G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler. Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems. *Multiscale Modeling & Simulation*, 7(2):842–864, 2008.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5): 603–619, 2002.

W. Czaja, B. Manning, L. McLean, and J.M. Murphy. Fusion of aerial gamma-ray survey and remote sensing data for a deeper understanding of radionuclide fate after radiological incidents: examples from the Fukushima Dai-Ichi response. *Journal of Radioanalytical and Nuclear Chemistry*, 307(3):2397–2401, 2016.

S. Dasgupta. Two faces of active learning. *Theoretical Computer Science*, 412(19): 1767–1781, 2011.

S. Dasgupta and L. Schulman. A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research*, 8(Feb):203–226, 2007.

F. Den Hollander. *Large Deviations*, volume 14. American Mathematical Soc., 2008.

P. Diaconis and L. Saloff-Coste. Comparison theorems for reversible Markov chains. *The Annals of Applied Probability*, pages 696–730, 1993.

P. Diaconis and D. Stroock. Geometric bounds for eigenvalues of Markov chains. *The Annals of Applied Probability*, pages 36–61, 1991.

E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (11):2765–2781, 2013.

M. Ester, H.-P. Kriegel, J. Sander, X. Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, volume 96, pages 226–231, 1996.

K. Fountoulakis, D.F., and M.W. Mahoney. An optimization approach to locally-biased graph algorithms. *Proceedings of the IEEE*, 105(2):256–272, 2017.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in Statistics Springer, Berlin, 2001.

K. Fukunaga and L. Hostetler. The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on Information Theory*, 21(1):32–40, 1975.

N Garcia Trillos, D. Slepcev, J. Von Brecht, T. Laurent, and X. Bresson. Consistency of Cheeger and ratio graph cuts. *Journal of Machine Learning Research*, 17(181): 1–46, 2016.

N. Garcia Trillos, M. Gerlach, M. Hein, and D. Slepcev. Error estimates for spectral convergence of the graph Laplacian on random geometric graphs towards the Laplace–Beltrami operator. *arXiv preprint arXiv:1801.10108*, 2018.

C. Gardiner. *Stochastic Methods*, volume 4. springer Berlin, 2009.

M. Gavish and B. Nadler. Normalized cuts are approximately inverse exit times. *SIAM Journal on Matrix Analysis and Applications*, 34(2):757–772, 2013.

C.R. Genovese, M. Perone-Pacifico, I. Verdinelli, and L. Wasserman. Non-parametric inference for density modes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):99–126, 2016.

J.C. Gower and G.J.S. Ross. Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 18 (1):54–64, 1969.

J.A. Hartigan. Consistency of single linkage for high-density clusters. *Journal of the American Statistical Society*, 76(374):388–394, 1981.

M. Jerrum and A. Sinclair. Approximating the permanent. *SIAM Journal on Computing*, 18(6):1149–1178, 1989.

S. Jia, G. Tang, J. Zhu, and Q. Li. A novel ranking-based clustering approach for hyperspectral band selection. *IEEE Transactions on Geoscience and Remote Sensing*, 54(1):88–102, 2016.

P.W. Jones, M. Maggioni, and R. Schul. Manifold parametrizations by eigenfunctions of the Laplacian and heat kernels. *Proceedings of the National Academy of Sciences*, 105(6):1803–1808, 2008.

S. Lafon, Y. Keller, and R.R. Coifman. Data fusion and multicue data matching by diffusion maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(11):1784–1797, 2006.

R.R. Lederman and R. Talmon. Learning the geometry of common latent variables using alternating-diffusion. *Applied and Computational Harmonic Analysis*, 44(3): 509–536, 2018.

R.R. Lederman, R. Talmon, H. Wu, Y.-L. Lo, and R.R. Coifman. Alternating diffusion for common manifold learning with application to sleep stage assessment. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5758–5762. IEEE, 2015.

D.A. Levin, Y. Peres, and E.L. Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

R. Li, M.G. Frasch, and H. Wu. Efficient fetal-maternal ECG signal separation from two channel maternal abdominal ECG via diffusion-based channel selection. *Frontiers in physiology*, 8:277, 2017.

S. Ling and T. Strohmer. Certifying global optimality of graph cuts via semidefinite relaxation: A performance guarantee for spectral clustering. *Foundations of Computational Mathematics*, pages 1–55, 2018.

A. Little and A. Byrd. A multiscale spectral method for learning number of clusters. In *International Conference on Machine Learning and Applications (ICMLA)*, pages 457–460. IEEE, 2015.

A. Little, M. Maggioni, and J.M. Murphy. Path-based spectral clustering: Guarantees, robustness to outliers, and fast algorithms. *ArXiv:1712.06206*, 2017.

M. Maggioni and J.M. Murphy. Learning by active nonlinear diffusion. *Foundations of Data Science*, 1(3):271–291, 2019.

M. Meila. Good (k-means) clusterings are unique (up to small perturbations). *Journal of Multivariate Analysis*, 173:1–17, 2019.

M. Meila and J. Shi. Learning segmentation by random walks. In *Advances in Neural Information Processing Systems (NIPS)*, pages 873–879, 2001.

C.D Meyer. Stochastic complementation, uncoupling Markov chains, and the theory of nearly reducible systems. *SIAM Review*, 31(2):240–272, 1989.

D.G. Mixon, S. Villar, and R. Ward. Clustering subgaussian mixtures by semidefinite programming. *Information and Inference: A Journal of the IMA*, 6(4):389–415, 2017.

J.M. Murphy and M. Maggioni. Diffusion geometric methods for fusion of remotely sensed data. In *Algorithms and Technologies for Multispectral, Hyperspectral, and Ultraspectral Imagery XXIV*, volume 10644, page 106440I. International Society for Optics and Photonics, 2018a.

J.M. Murphy and M. Maggioni. Iterative active learning with diffusion geometry for hyperspectral images. In *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5. IEEE, 2018b.

J.M. Murphy and M. Maggioni. Spectral-spatial diffusion geometry for hyperspectral image clustering. *arXiv preprint arXiv:1902.05402*, 2019a.

J.M. Murphy and M. Maggioni. Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3):1829–1845, 2019b.

B. Nadler and M. Galun. Fundamental limitations of spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1017–1024, 2007.

A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems*, 2:849–856, 2002.

R. Ostrovsky, Y. Rabani, L.J. Schulman, and C. Swamy. The effectiveness of Lloyd-type methods for the k-means problem. In *Symposium on Foundations of Computer Science (FOCS)*, pages 165–176. IEEE, 2006.

H.-S. Park and C.-H. Jun. A simple and fast algorithm for k-medoids clustering. *Expert Systems with Applications*, 36(2):3336–3341, 2009.

A. Rodriguez and A. Laio. Clustering by fast search and find of density peaks. *Science*, 344(6191):1492–1496, 2014.

M.A. Rohrdanz, W. Zheng, M. Maggioni, and C. Clementi. Determination of reaction coordinates via locally scaled diffusion map. *The Journal of Chemical Physics*, 134 (12):03B624, 2011.

C. Rossant, S.N. Kadir, D.F.M. Goodman, J. Schulman, M.L.D. Hunter, A.N. Saleem, A. Grosmark, M. Belluscio, G.H. Denfield, A.S. Ecker, et al. Spike sorting for large, dense electrode arrays. *Nature Neuroscience*, 19(4):634, 2016.

S.T. Roweis and L.K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

G. Schiebinger, M.J. Wainwright, and B. Yu. The geometry of kernelized spectral clustering. *The Annals of Statistics*, 43(2):819–846, 2015.

J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

T. Shi, M. Belkin, and B. Yu. Data spectroscopy: Eigenspaces of convolution operators and clustering. *The Annals of Statistics*, pages 3960–3984, 2009.

H.A. Simon and A. Ando. Aggregation of variables in dynamic systems. *Econometrica: Journal of the Econometric Society*, pages 111–138, 1961.

A. Sinclair and M. Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, 1989.

A. Singer and R.R. Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

A. Singer and H. Wu. Vector diffusion maps and the connection Laplacian. *Communications on Pure and Applied Mathematics*, 65(8):1067–1144, 2012.

A. Singer and H. Wu. Spectral convergence of the connection Laplacian from random samples. *Information and Inference: A Journal of the IMA*, 6(1):58–123, 2016.

A. Singer, R. Erban, I.G. Kevrekidis, and R.R. Coifman. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proceedings of the National Academy of Sciences*, 106(38):16090–16095, 2009.

P.H.A. Sneath. The application of computers to taxonomy. *Microbiology*, 17(1): 201–226, 1957.

M. Soltanolkotabi, E. Elhamifar, and E.J. Candes. Robust subspace clustering. *Annals of Statistics*, 42(2):669–699, 2014.

D.A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Symposium on Theory of Computing (STOC)*, pages 81–90. ACM, 2004.

D.A. Spielman and S.-H. Teng. A local clustering algorithm for massive graphs and its application to nearly linear time graph partitioning. *SIAM Journal on Computing*, 42(1):1–26, 2013.

D.A. Spielman and S.-H. Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM Journal on Matrix Analysis and Applications*, 35(3):835–885, 2014.

M.H. Spitzer, P.F. Gherardini, G.K. Fragiadakis, N. Bhattacharya, R.T. Yuan, A.N. Hotson, R. Finck, Y. Carmi, E.R. Zunder, W.J. Fantl, et al. An interactive reference framework for modeling a dynamic immune system. *Science*, 349(6244):1259425, 2015.

H. Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Polon. Sci.*, 4(12):801–804, 1957.

K. Sun, X. Geng, and L. Ji. Exemplar component analysis: A fast band selection method for hyperspectral imagery. *IEEE Geoscience and Remote Sensing Letters*, 12(5):998–1002, 2015.

G. Szegö. Inequalities for certain eigenvalues of a membrane of given area. *Journal of Rational Mechanics and Analysis*, 3:343–356, 1954.

R. Talmon and H. Wu. Latent common manifold learning with alternating diffusion: analysis and applications. *Applied and Computational Harmonic Analysis*, 2018.

J.B. Tenenbaum, V. De Silva, and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

R. Vidal. Subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2011.

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

Y. Wan and M. Meila. Graph clustering: block-models and model free results. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2478–2486, 2016.

P. Wang, B. Xu, J. Xu, G. Tian, C.-L. Liu, and H. Hao. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification. *Neurocomputing*, 174:806–814, 2016.

X. Wang, K. Slavakis, and G. Lerman. Multi-manifold modeling in non-Euclidean spaces. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 1023–1032, 2015.

H.F. Weinberger. An isoperimetric inequality for the n-dimensional free membrane problem. *Journal of Rational Mechanics and Analysis*, 5(4):633–636, 1956.

C. Wiwie, J. Baumbach, and R. Röttger. Comparing the performance of biomedical clustering methods. *Nature Methods*, 12(11):1033, 2015.

X. Xu, M. Ester, H.-P. Kriegel, and J. Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *International Conference on Data Engineering (ICDE)*, pages 324–331. IEEE, 1998.

H. Yin, A.R. Benson, J. Leskovec, and D.F. Gleich. Local higher-order graph clustering. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 555–564. ACM, 2017.

L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1601–1608, 2005.

T. Zhang, A. Szlam, Y. Wang, and G. Lerman. Hybrid linear modeling via local best-fit flats. *International Journal of Computer Vision*, 100(3):217–240, 2012.

W. Zheng, M.A. Rohrdanz, M. Maggioni, and C. Clementi. Polymer reversal rate calculated via locally scaled diffusion map. *The Journal of Chemical Physics*, 134 (14):144109, 2011.