POLYTOPE KERNEL DENSITY ESTIMATES ON DELAUNAY GRAPHS

Erhan Bas, Deniz Erdogmus

Cognitive Systems Laboratory, ECE Department, Northeastern University, Boston, MA 02115, USA

ABSTRACT

We present a polytope-kernel density estimation (PKDE) methodology that allows us to perform exact mean-shift updates along the edges of the Delaunay graph of the data. We discuss explicit and implicit constructions of such a PKDE, where in the implicit construction one can exploit a smoother kernel such as the standard isotropic Gaussian. The resulting density estimate allows us to perform mean-shift clustering in a computationally efficient manner (similar to mediod shift), but in a manner that is exact and consistent with the underlying density assumption. The procedure also yields a hierarchical connectivity structure, a tree, that spans the dataset. We demonstrate how this tree, combined with density-weighted geodesic distance calculations between modal samples can be used to select number of clusters as well as a distance preserving dimension reduction technique.

Index Terms— Polytope kernel density estimation, mean shift clustering

1. INTRODUCTION

Manifold learning has been extensively studied [1, 2, 3, 4] and the fundamental underlying mathematical problem has reincarnated many times over the last four decades in the form of least orthogonal error least squares regression, modeling with errors-in-variables, principal surfaces, and nonlinear principal components. The goal is to determine a representation of a randomly distributed possibly high dimensional data with a probability distribution that is tightly concentrated on a low-dimensional (piecewise) smooth surface, the so-called underlying manifold of the data distribution. The solution to this problem can be utilized in tackling many fundamental statistical inference and machine learning problems including clustering, dimensionality reduction, signal denoising, and nonlinear warping for registration or coordinate alignment. Typical methods for determining the optimal manifold utilize minimum mean-squared-error (MSE) type objective functions. For instance, linear principal component analysis (PCA) yields a minimum-MSE hyperplane as the optimal underlying manifold, which makes geometrical sense when the underlying data distribution is elliptically symmetric, such as a Gaussian with anisotropic covariance. In general, a reasonable parametric model family is difficult to select for data distributions that are complicated, especially for high dimensional data; consequently, techniques that focus on nonparametric techniques such as kernel machines, or kernel density estimation, as well as those that utilize neighborhood graph constraints (which are similar in spirit to our proposal in this paper) have been preferred and achieved successful results in clustering and dimension reduction.

We note that not all densities will have globally smooth underlying manifolds that can be nicely unwrapped or projected onto a corresponding Euclidean space; consider for instance a mixture of two elongated Gaussians that are positioned and oriented to form a 'T' shape in 2-dimensional space - intuitively the best 1-dimensional approximation data, as one would obtain, for instance using local PCA, consists of two separate (approximately) linear 1-dimensional segments [4]. Consequently, we assert that underlying low dimensional manifolds for an arbitrary data distribution in general exhibit a segmented piecewise nonlinear and smooth structure that can be extracted from a graph (or in some cases, as a very good approximation, a tree structure). In this paper, along this line of reasoning, which is illustrated in our earlier work [4], we propose a local to global tree-structured topology extraction technique based on a polytope kernel density estimate (PKDE) framework that establishes pairwise data connections that are significant in terms of the geometry of the estimated data distribution, not just Euclidean distances between data pairs. The formulation gives rise to a hierarchical local cluster representation that can be obtained in a computationally efficient manner similar to mediod-shift [5]; in fact, we claim that the PKDE formulation leads to a rigorous derivation of mediod-shift as a clustering technique as isotropic Gaussian-KDE (GKDE) leads to mean-shift (MS) clustering [6, 7]. The local clusters are connected through a minimum spanning tree that utilizes geodesics between modes (central points that represent the cluster peaks).

Therefore, the contributions of this paper are two-fold: (i) we introduce PKDE as a methodology to approximately extract underlying cluster and manifold structure, (ii) and we present a fast MS algorithm based on PKDE (which leads to piecewise linear density approximations) and linear programming. We leave extensive treatment of how to use the PKDE for manifold learning with higher dimensions to future work for lack of space.

This work is supported by NSF under grants ECCS0929576, ECCS0934506, IIS0934509, IIS0914808, and BCS1027724. The opinions presented here are solely those of the authors and do not necessarily reflect the opinions of the funding agency.

2. POLYTOPE KERNEL DENSITY ESTIMATION

MS is a popular and successful clustering technique that suffers from high computational complexity; various simplifications have been investigated including finite-support kernels and space discretization [6, 8, 7, 9, 5, 10]. As opposed to existing techniques that typically start from a continuously differentiable KDE (such as a GKDE), we propose an MS variant that assumes a PKDE based on the use of finite-support polytope-shaped kernels (basically pyramids with convex polygon bases in data space) whose supports are determined by the Delaunay graph that spans the data points. Artificial edges resulting from boundary data points are eliminated by deleting the edges that are on the convex-hull boundary of the whole dataset. This process partitions the data space into simplexes in which the probability density is approximated as a linear surface; therefore a hill-climbing procedure for each data point can be obtained by solving multiple linear programs (linear density to be maximized within a simplex-shaped feasible set) and then selecting the best solution for each data point (vertices of the simplexes) across all simplexes it belongs to. This reduces to connecting each data point to its highest density neighbor in the Delaunay graph. This process yields a clustering solution and a tree-structured hill-climbing connectivity map that spans each cluster. A global cluster spanning tree that connects the modes using paths along the Delaunay graph is then obtained using a density-geodesic concept [11].

For the described linear program based MS clustering approach to work, we need a piecewise-linear KDE defined on simplexes forming a partition of the data support (e.g. convex hull). Such a KDE can be obtained by employing polytope kernels. While in general data points do not have to be vertices of the simplexes and polytope supports, ϵ-ball graphs have been found to be useful in practice when using Euclidean distances; consequently, we employ this using graph-geodesic distances over Delaunay graphs of the data. The Delaunay graph is obtained by finding the Voronoi partition of the data space and connecting data (nodes) whose Voronoi cells share a boundary with an edge. In order to eliminate possible inter-cluster edges that could form between samples of clusters with different scales, we delete edges connecting a data point to its neighbors if this action improves the *uniformity* of edge lengths (measured by using entropy of edge lengths after normalization to unit sum and treating them as probability masses).

We propose to utilize variable-width polytope kernels; a polytope-kernel centered at a data point $\mathbf{x} \in \mathbb{R}^n$ is a pdf that has bounded support on an n-dimensional polytope whose vertices, edges, and (hyper-) surfaces are defined by the set of points $\mathbf{y}^{\mathbf{i}}_{\mathbf{x}}, i = 1, ..., K$ that are within $\epsilon \in \mathbb{Z}$ distance from $\mathbf{x}$ on the Delaunay graph and the graph edges that connect these points to each other (i.e. the convex hull of these points). The neighborhoods are obtained by employing a shortest path algorithm [12] on the graph assuming unit edge lengths. The kernel, following usual convention, is constructed to have its peak value at $\mathbf{x}$ and has linear *faces* on simplexes that connect $\mathbf{x}$ to each subset of n points in the ϵ-ball set defined above. Computational details and equations for constructing these kernels will be included in the journal extension of paper due to lack of space. This process describes the explicit construction of polytope kernels leading to a PKDE when used as usual, setting ϵ to 3 or 5, for instance.

Alternatively, in an implicit polytope kernel selection approach, one could construct the Delaunay graph as well as smooth density estimate (for instance a GMM or a KDE using a smooth kernel such as Gaussian). Then the smooth density can be sampled at the data points and for each point, for each data point, the density could be linearly approximated within each simplex formed by a given data point and n of its adjacent neighbors in the graph (doing this for every n-element neighbor subset) by determining the linear function that satisfies the sample values at the data points. Interior of the simplex in question can be spanned by a convex linear combination of its vertices and the linear approximation is given by the same weighted linear combination of sampled smooth KDE values at the vertices. The implicit method has two apparent advantages: (1) it approximates a smooth KDE in a piece-wise fashion so spurious peaks are less likely to emerge (important for MS clustering), (2) these are easier to understand intuitively - although for each implicit PKDE there is a corresponding polytope kernel selection process and one could have obtained the same result following the explicit procedure described above. Fig. 1 shows explicit and implicit PKDE models for a Gaussian sample.

3. CLUSTERING BY LINEAR PROGRAMS

Given a PKDE, we obtain a density model that is piecewise linear on simplexes whose vertices are data points, if one of the two strategies mentioned above is used. If a simplex S is defined by $n+1$ data points in n–dimensions, the density for a point $\mathbf{x} \in S \subset \mathbb{R}^d$ is given by $p(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$. For a given simplex, the linear program ($\arg_{\mathbf{x}} \max p(\mathbf{x})$ subject to $\mathbf{x} \in S$) finds the vertex with the highest density value; consequently, for all other vertices, which are data points, a good hill-climbing candidate is given by the adjacent data point on the simplex with largest value. In practice we don't need to solve linear programs; the MS update simply becomes finding for each data point, among its adjacent neighbors on the Delaunay graph constructed as described above, the one that has the largest probability density value according to PKDE. Most of the time, we expect that the maximal value in each linear program is achieved by only one vertex of S. Two possible problematic cases are: (1) $\mathbf{w}^T\mathbf{e} = 0$, where $\mathbf{e}$ is the vector parallel to the edge on which the maximum density is achieved - two vertices on this edge are both possible choices for the iteration; (2) $\mathbf{w} = \mathbf{0}$, which yields a constant value of density within the simplex - thus all vertices are possible solutions. Other problematic cases between these two extremes

exist. One potential remedy is to chose the iteration for the current data in question such that the selected iterated vertex then iterates itself to a larger density value.

This procedure creates a hierarchical tree structure within each mode where each data is connected to a parent data via the edge that it follows while solving the LP problems in its vicinity. Consequently, the algorithm requires only one iteration per sample - and that iteration is quite simple: go to the neighbor with the largest density. Similar ideas have been explored in the MS literature [9, 5, 10]. Unlike previous methods, since our approach utilizes piecewise linear surfaces as density estimates, updates emerging from the formulated problem are not approximates for gradient but exact MS iterations constrained to the given graph. Clearly, the root node of each cluster's tree will naturally be the data with the highest density in that cluster. Furthermore, since the structure is a tree, one can utilize the tree to constrain the pairwise distances to be maintained in dimension reduction approaches such as LLE or ISOMAP [3, 1] in order to obtain a two-dimensional projection of the data for visualization or compression purposes (since a tree is a planar graph). The same will be true for the general tree structure that we fit to the data globally; details will be in the next section. Specifically, we will try to learn the global topology of clusters and their principal curves using shortest paths between modes and minimum spanning trees that traverse the modes of each cluster.

4. MINIMUM SPANNING TREES TO MERGE CLUSTER INDUCED CHARTS INTO AN ATLAS

MS clustering is known to yield over-segmentation results if the KDE generates many peaks. Mode-merging is a successful modal order control technique if used properly. Intuitively, if the ridge (principal curve according to our definition [4] connecting two modes do not drop in value too much (minimum on the ridge occurs on the saddle point for a smooth KDE), then the modes could be merged into a single cluster since, very likely, the presence of two modes as opposed to one is a statistical anomaly due to finite sample size. For PKDE, first and second order derivatives are not well defined (generalizations exist for nonsmooth functions and we will investigate these in the future); consequently, we employ a inverse-density-weighted geodesic distance approach to decide whether two modes should be merged or not [11].

In particular, the weight of each edge in the Delaunay graph is taken as a monotonic function of the generalized mean of the density along the edge multiplied by the Euclidean length of the edge in the data space: $d = ||x_i - x_j|| h^{-1} \int_{edge} h(p(\mathbf{x}))$. Specifically, we utilize the function $h(a) = 1/a$, which corresponds to the harmonic mean, but other choices are possible. For the case of linear edges and linear densities along the edge, this *distance* between data samples i and j can be calculated analytically:

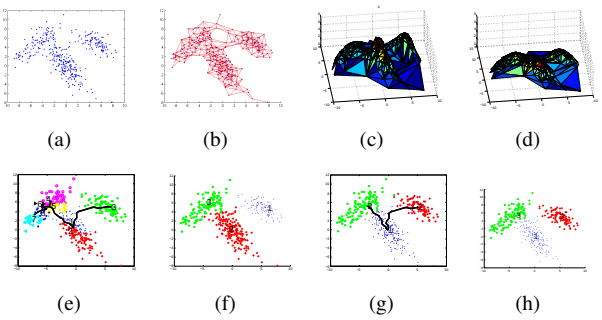


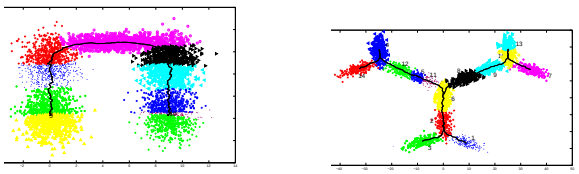
Fig. 1. Synthetic dataset composed of mixture of 3 Gaussians: (a) GMM samples, (b) Delaunay graph after deleting boundary and intra-cluster edges having low entropy values, (c) KDE using polytope kernels, (d) Piecewise linear KDE approximation using samples from a Gaussian-KDE, (e) Initial clustering with polytope-KDE, (f) Merging of clusters in (e) using mode connectivity, (g) Initial clustering with Gaussian KDE approximation, (h) Gaussian-KDE mean-shift clustering. Selection of neighborhood radius and kernel width will effect the density estimate. We used a graph neighborhood radius of 4 for polytope kernels and 1.5 for isotropic Gaussian kernel-width.

$$\begin{aligned} d_{ij} &= ||x_i - x_j|| \int_0^1 \frac{1}{p(x(t))} dt \\ &= ||x_i - x_j|| \frac{1}{p(x_j) - p(x_i)} ln(\frac{p(x_j)}{p(x_i)}). \end{aligned} \quad (1)$$

As a result, the density-weighted geodesic distance between two cluster modes can be given as $\sum_{edge_{ij} \in C} d_{ij}$ where C is the shortest path between the modes calculated on the edge graph with weights as the length penalized harmonic density averages [12].[1] The pairwise mode distances obtained above form a fully connected pairwise distance graph between the modes, which are subjected to a Minimum Spanning Tree (MST) search algorithm [13] in order to obtain a tree-structured sparse connectivity graph between the modes, thus creating a planar (but one-dimensional in local structure) global atlas of coordinates for the data.

Fig. 1(e-f) show clustering results obtained using explicit and implicit polytope kernels (from GKDE in the latter). The modes are connected via MST as described above using the harmonic density average-weighted edge lengths in (1). Each color represents a cluster with black curves representing the tree that spans the cluster modes. Overall, the MST that connects the modes globally, and then the hierarchical cluster trees within each modal cluster reveal a global tree structure for the data, and since trees are planar graphs, this tree could be used to reduce the dimensionality of the data to 2 from n using an algorithm like LLE where the local pairwise distances to be preserved during optimization are limited to the neighbor pairs in the data tree; thus the low dimensional data maintains the same tree structure as well. In this paper, we do

[1] One could find the shortest *harmonic* distance instead of *arithmetic*.



(a) $\prod$ data (b) Branch data

Fig. 2. Results on synthetic datasets.

(a) (b) (c) (d)

Fig. 3. Image segmentation results: (a) Original image, (b) 5-cluster output using the proposed algorithm, (c) 6-cluster output using the proposed algorithm, (d) Corresponding Gaussian-KDE mean-shift result with 6 clusters.

not investigate this extension further.

5. RESULTS

We illustrate clustering results on two other synthetic Gaussian mixtures. Π-shaped dataset consists of 5 cascaded elongated Gaussians. Each component has 1000 samples with diagonal covariance with eigenspread 4. Fig. 2(a) illustrates the algorithm output for this dataset. Second synthetic dataset consists of a Gaussian mixture forming a tree with 9 branches connected at four 3-way bifurcation points. Middle branches consist of two Gaussian components making the mixture have 12 components where each one has 200 samples with an eigenspread of 4 again. The underlying pdf is estimated with GMM using expectation-maximization, where the number of components is selected as the number of actual components. Fig. 2(b) illustrates the algorithm result on this dataset.

We applied the proposed method to grayscale image clustering using pixel coordinate and intensity feature vectors. The benchmark cameraman image, with 50x50 size is used and segmentation results are shown in Fig. 3. Kernel width of the Gaussian kernel density estimate is manually selected to obtain the desired number of clusters. In general proposed method results in more consistent clusters with less number of outliers, i.e. top of chimney, compared to GKDE. However, since selected polytope kernels have arbitrary shapes, cluster boundaries might have non-smooth shapes around homogenous regions, e.g. vertical sky boundary.

6. DISCUSSION

In this paper we proposed using a polytope kernel density estimate (PKDE) for density modeling and illustrated its application to mean-shift clustering, leading to a simple clustering algorithm that uses exact MS updates under the assumed kernel type while, as opposed to alternatives in the literature that are approximate with respect to their models. We also demonstrated how mode merging can be achieved using a density-weighted geodesic distance on the Delaunay graph, leading to a spanning tree for the data, which highlights the clustering structure in a hierarchical manner. This tree could be employed in dimension reduction and manifold learning as a constraint. Main practical contributions are: *i*) a computationally efficient MS clustering algorithm that requires only one simple update per sample; *ii*) a spanning tree structure for local clusters as well as cluster modes that provides a hierarchical spanning graph representation that could be used as a constraint for dimension reduction that attempts to maintain high dimensional distances and neighborhood structures; *iii*) if the spanning tree for cluster modes is replaced by a pruned graph with possible loops, bifurcations and self intersections in the underlying graph can be identified using geodesic paths between modes.

7. REFERENCES

[1] J. B. Tenenbaum, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, December 2000.

[2] Pierre Gaillard, Michaël Aupetit, and Gérard Govaert, "Learning topology of a labeled data set with the supervised generative gaussian graph," *Neurocomputing*, vol. 71, no. 7-9, pp. 1283–1299, 2008.

[3] Lawrence K. Saul, Sam T. Roweis, and Yoram Singer, "Think globally, fit locally: Unsupervised learning of low dimensional manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, 2003.

[4] Umut Ozertem and Deniz Erdogmus, "Local conditions for critical and principal manifolds," in *Proceesings of ICASSP 2008*, pp. 1893–1896.

[5] Yaser Ajmal Sheikh, E.Khan, and Takeo Kanade, "Mode-seeking by medoidshifts," in *ICCV 2007*, Oct 2007.

[6] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. Information Theory*, vol. 21, no. 1, pp. 32–40, Jan 1975.

[7] Yizong Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790–799, Aug 1995.

[8] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 5, pp. 603–619, May 2002.

[9] M.A. Carreira-Perpinan, "Acceleration strategies for gaussian mean-shift image segmentation," in *Proceedings of CVPR 2006*, June 2006, vol. 1, pp. 1160–1167.

[10] A. Vedaldi and S. Soatto, "Quick shift and kernel methods for mode seeking," *ECCV 2008*, pp. 705–718, 2008.

[11] Umut Ozertem, Deniz Erdogmus, and Miguel A. Carreira-Perpinan, "Density geodesics for similarity clustering," in *Proceesings of ICASSP 2008*, March 2008, pp. 1977–1980.

[12] Erwin Kreyszig, *Advanced Engineering Mathematics*, John Wiley & Sons, Inc., New York, NY, USA, 1972.

[13] R.L. Rivest T.H. Cormen, C.E. Leiserson, *Introduction to Algorithms*, MIT Press & Mc Graw-Hill, New York, USA, 1990.