Parameter-Free Density Estimation for Hyperspectral Image Clustering

Steven Le Moan, Claude Cariou

***Abstract*—This paper investigates multivariate kernel density estimation for hyperspectral data. More specifically, it focuses on designing a data-driven, parameter-free and fast estimator that is able to deal with clusters of different size, shape and density. We first propose a new criterion to compare density estimates, based on how evenly they treat the different classes in the data. In addition, we propose a simple, parameter-free and non-iterative kernel bandwidth selection approach. We show that it is substantially more adaptive than existing estimators on six remote sensing hyperspectral datasets.**

***Index Terms*—Clustering, Hyperspectral, Density, Parameter-free.**

I. Introduction

Hyperspectral imaging produces large, high-dimensional and complex datasets from which it is often challenging to extract information. Reference data, known as *ground truth*, is typically used in order to facilitate their analysis in so-called *supervised* methods. Unfortunately, this ground truth is difficult to obtain in remote sensing applications, as it involves surveying campaigns that are long, expensive, as well as error-prone (see e.g. [1]). In view of this, *unsupervised* methods are very desirable in that they need no reference data.

This paper investigates unsupervised pixel classification, also known as *clustering*. In an image, pixels can be grouped, or *clustered*, based on what they represent, like for example a certain type of vegetation, mineral, etc. The purpose of clustering is to identify these groups from the pixel values only. This is done in the feature space spanned by the spectral bands, where data points (pixels) form clusters. These can be defined based on different paradigms such as contiguity [2], distance to a centroid [3] or density [4]. Hyperspectral images require methods that are able to recognise clusters of **arbitrary shape** (e.g. not necessarily convex) and **size** in a **high dimensional** space, and in the presence of **noise**. To that end, density-based methods have proven particularly useful. They work under the assumption that hyperspectral pixels are sampled from a multivariate probability density function of unknown characteristics, which can be estimated from the data. Clusters are then defined as regions stemming from local maxima of that function.

In hyperspectral images, it is common to encounter clusters of arbitrary shape, size, but also **sparsity**. A cluster with high sparsity typically represents a class of material with a larger dynamic range of reflectance values. It will occupy a larger region of the feature space than other clusters, and the average pairwise distance between its members will also be larger. Examples of clusters with different sparsities are shown in Fig. 1. From a signal processing perspective, a sparse cluster can be seen as a discrete signal that is sampled from its underlying probability density function at a low sampling rate. The presence of clusters of mixed sparsity in a dataset then either means that the pixels are non-uniformly sampled from a single multi-modal probability density function (pdf) $d(\mathbf{x})$, or that pixels in each cluster are sampled uniformly from an individual pdf. In the latter case, $d(\mathbf{x})$ is the normalised sum of all individual pdfs.

This view is opposed to the traditional assumption that the data is uniformly sampled from a single pdf, with peaks of

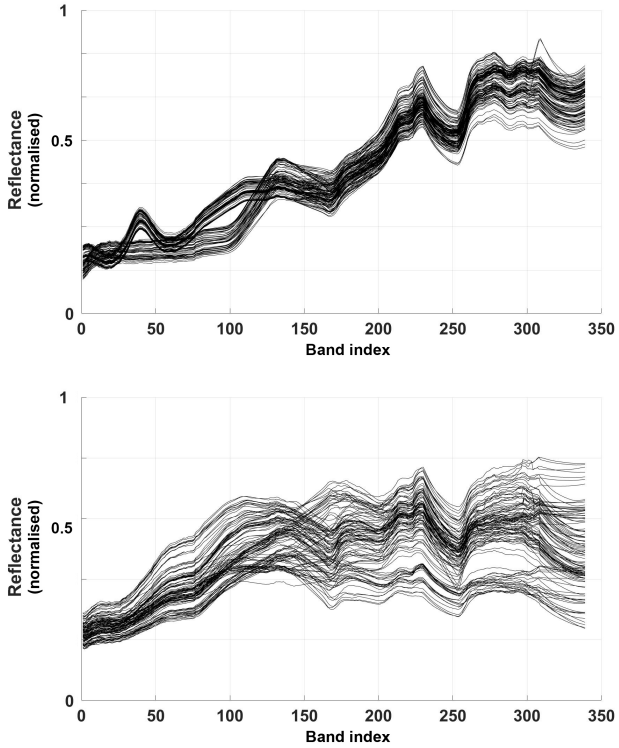

Figure 1. Example of high-dimensional clusters with low (top) and high (bottom) sparsity from the Massey University dataset (see Section IV-B). Note that the top cluster corresponds to a single type of vegetation, which can be further decomposed into two sub-clusters based on their reflectance in the visible range of wavelengths.

Steven Le Moan is with the School of Engineering and Advanced Technology at Massey University, Palmerston North, New Zealand (e-mail: s.lemoan@massey.ac.nz). Claude Cariou is with the Institute of Electronics and Telecommunications of Rennes, University of Rennes 1 - Enssat - 6 rue de Kerampont, 22300 Lannion, France

978-1-7281-0125-5/18/$31.00 ©2018 IEEE

different magnitude in each cluster. Most existing approaches such as the very popular Mean Shift [5] or the Fast Density Peak Clustering (FDPC) [6] implicitly rely on this assumption and then calculate local maxima of this function to identify *modes*, *ridges* or *core points* in the data [5]–[10]. A correct identification of modes permits to determine the number of clusters in the data, but does not guarantee an adequate density estimate. As pointed out in numerous studies, labelling low density data points is more ambiguity-prone as they do not represent well their respective clusters. It is then important that data points are properly ranked in terms of density so as to identify core points, which can in turn be used for a more robust clustering. The result of that clustering is then propagated to non-core points e.g. via a nearest-neighbour rule [11]. Other methods rely on processing data points (pixels) in order of decreasing density [6], [12] or threshold the density function to make the clustering more robust [11]. It is important not only that the density method concentrates its energy in compact regions that characterise cluster structures, but also that it distributes its energy somewhat evenly to all clusters. In other words, each cluster must be given the same importance so that the set of $x\%$ points of highest density contain representative proportions of each cluster, regardless of their sparsity and for any x. This is the rationale behind the criterion we propose in this paper, which we call the Degree of Imbalance (DI) of the density estimator.

Furthermore, most existing density estimation methods are burdened by parameter tuning, which greatly influences their results. In FDPC, the density of a point p is defined as the number of other points lying within a hyper-sphere centered at p of radius d_c, a parameter of the method. The recently proposed method GWENN [12] uses the kNN principle to label points in order of decreasing density and defines a mode as a point with none of its kNN already labelled. In using the watershed principle, GWENN starts with the least ambiguous cases, which makes it less sensitive to noise in low-density points. The method also features a wavelet-based multi-scale scheme which permits the clustering of large images. However, GWENN and other kNN-based methods [13], [14] depend heavily on parameter tuning, despite their relative independence to scale, dimensionality and sparsity [15]. Rules of thumbs are often used to determine the optimal k (e.g. 5% of the size of the data) but there is currently no efficient and accurate approach. The *elbow* method (see e.g. [12]) is imprecise, noise-sensitive but also iterative, which implies a significant computational cost. One of the most popular classes of density estimators are those that are kernel-based [16]. They are simple and fast, but also rely on tuning a bandwidth. Kernel density estimation is used for example in Mean Shift [5], and has been successfully applied to the analysis of high dimensional data [17]. Tuning the kernel's bandwidth can be done via a rule of thumb [18], or from the data [19]–[21] but often to the expense of computational cost.

In addition to the DI criterion, we also propose a new automatic bandwidth selection method that is optimal in the sense of the Nyquist criterion and under the assumption that the data is under-sampled from its underlying pdf. We demonstrate that the proposed method significantly outperforms other parameter-free density estimators in terms of DI and that it can effectively concentrate its energy on the underlying manifold of clusters in high dimensionality.

We use the following notation in the remainder of this paper:

- $\mathbf{x} = \{x_1, \ldots, x_D\}$ is a D-dimensional random variable representing a hyperspectral pixel.
- $d(\mathbf{x})$ is the multivariate pdf of $\mathbf{x}$, and $\hat{d}(\mathbf{x})$ is its estimator.
- $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is the dataset, made of N pixels.
- $\delta_{i;j}$ is the distance between $\mathbf{x}_i$ and its j-th neighbour.

II. Degree of Imbalance

Correctly identifying density modes and ridges is central to density-based clustering. This is particularly true for those approaches that rely on ranking the data points in terms of density, whereby modes are used to initialise the clustering. A wrong initialisation can lead to significant error propagation in the rest of the data. Typically, clustering accuracy is used as a criteria to validate density estimators, but gives little information about the actual relevance of the density estimators itself. Here, we propose a new external (i.e. supervised) criterion that measures how evenly the density estimator distributes its energy to the different clusters in the data.

Let us consider that we want at least $\tau\%$ of each class represented in our set of core points. Following our rationale, taking the $\tau\%$ most dense pixels (according to $\hat{d}(\mathbf{x})$) should systematically result in a set where each class is represented by at least $\tau\%$ of its total population. For instance, if we have two classes C1 and C2, of sizes 10 and 100 respectively, the set of 10% most dense pixels should contain 1 pixel from C1 and 10 from C2 in order to be truly representative of the data. On the other hand, the set of 1% most dense pixels should contain 1 pixel from C1 and 1 from C2, as each class needs to be represented for all values of τ. As demonstrated in Section IV, most existing parameter-free density estimation methods result in some clusters being not represented at all in the set of $\tau\%$ most dense pixels, particularly for small values of τ.

We then note $\Delta(\tau)$ the size (as a percentage, normalised to N) of the smallest set of most dense pixels that contains at least $\tau\%$ of each class. Ideally, we should have $\Delta(\tau) = \tau$. If $\Delta(\tau) > \tau$, this means that some clusters are given more importance than others based on their size and sparsity, which makes it difficult to find a representative set of core points by thresholding $\hat{d}(\mathbf{x})$. This situation should be avoided, particularly for methods that proceed to label pixels in order of decreasing density. In order to derive a single-value score representing the local adaptability of a density estimator, we proposed to define the DI criterion as the 2-norm:

$$\mathrm{DI}(\hat{d}) = \|\Delta(\tau) \quad \tau\|_2 \,. \tag{1}$$

Therefore, the Degree of Imbalance is always positive or null, and it should be as small as possible.

III. Automatic bandwidth selection for kernel density estimation based on the Nyquist frequency

In addition to a small DI, our objectives in terms of density estimation are: speed and robust mode detection. We propose a simple parameter-free bandwidth tuning method that performs well in terms of these three criteria, especially in terms of DI.

The Nyquist frequency is the minimal frequency at which a continuous signal needs to be sampled to make sure it can be reconstructed without ambiguity. It is equal to twice the bandwidth of the continuous signal. This means that a discrete signal with sample frequency f_s permits reconstructing the underlying continuous signal without ambiguity only at frequencies up to $\frac{f_s}{2}$. Assume that $d(\mathbf{x})$ is one-dimensional and band-limited with a maximal frequency[1] f_0, it can only be recovered fully from data sampled at a rate greater than $2f_0$. This extends to the multivariate case by considering a different bandwidth in each dimension. However, for the sake of simplicity, we normalised all our experimental data to zero mean and unit variance. This allows us to use of an isotropic kernel with the same bandwidth in each dimension, that is substantially more computationally efficient than an anisotropic kernel. The type of kernel function is known to be of little significance in the general case [22] so we chose the traditional Gaussian kernel for our experiments.

In hyperspectral images, pixels are typically under-sampled from $d(\mathbf{x})$, in part due to the fact that high-dimensional spaces are mostly sparse and require a tremendous number of data points to be significantly populated. We then consider $\frac{f_s}{2}$ a data-driven upper bound for the cutoff frequency σ_f of any smoothing kernel $\mathbf{K}$ used for density estimation.

Sampling from a probability distribution typically results in spatial *non-uniformity* of samples, which then brings the question of determining the sampling rate f_s. It can be estimated globally, as the mean distance to a point's first neighbour, in which case the sampling frequency is given by:

$$\hat{f}_{s;g} = \frac{1}{N} \sum_{i=1}^{N} \delta_{i;1}. \tag{2}$$

Alternatively, it can be estimated locally as the distance to each point's first neighbour, in which case the sampling frequency is given pixelwise by:

$$\hat{f}_{s;l}(\mathbf{x}_i) = \delta_{i;1}. \tag{3}$$

We then define the kernel's cutoff frequency:

$$\sigma_f = \frac{1}{2\hat{f}_s}. \tag{4}$$

where $\hat{f}_s = \hat{f}_{s;g}$ or $\hat{f}_s = \hat{f}_{s;l}$.

From the definition of the D-dimensional Fourier transform, it comes that the frequency and spatial bandwidths (noted here σ_f and σ, respectively) of a multivariate Gaussian function are related through:

$$\sigma = \frac{1}{2\pi\sigma_f}, \tag{5}$$

which gives, in the case of the global estimator:

$$\sigma_g = \frac{1}{2\pi\frac{1}{2\hat{f}_{s;g}}} = \frac{\frac{1}{N}\sum_{i=1}^{N}\delta_{i;1}}{\pi} \tag{6}$$

where σ_g represents the globally estimated spatial bandwidth. In the case of the local estimator, we have:

$$\sigma_l(\mathbf{x}_i) = \frac{1}{2\pi\frac{1}{2\hat{f}_{s;l}}} = \frac{\delta_{i;1}}{\pi} \tag{7}$$

Ultimately, these bandwidths are used in a Gaussian kernel-based density estimator. The proposed *global* estimator is:

$$\hat{d}_g(\mathbf{x}_i) = \frac{1}{\sigma_g\sqrt{2\pi}^D} \sum_{j=1}^{N} e^{-\frac{\delta_{i;j}^2}{2\sigma_g^2}}, \tag{8}$$

and the proposed *local* estimator:

$$\hat{d}(\mathbf{x}_i) = \frac{1}{\sigma\sqrt{2\pi}^D} \sum_{j=1}^{N} e^{-\frac{\delta_{i;j}^2}{2\sigma^2}}. \tag{9}$$

The proposed density estimators are then optimal in the sense of the Nyquist criterion and under the assumption that the data is under-sampled from its underlying pdf.

IV. Experimental Validation

A. Parameter-free density estimators

The following global and local parameter-free density estimators were considered for comparison. For the kernel estimator, we used a Gaussian kernel, keeping in mind that the choice of the kernel's function is not as important as the selection of its bandwidth. As previously mentioned, for the sake of computational cost, we used only isotropic kernels, i.e. with the same bandwidth in each direction.

1) $\hat{d}_l(\mathbf{x}_i)$: isotropic Gaussian kernel with a *varying* bandwidth equal to the distance to the first nearest neighbour of each point, normalised by π (proposed estimator - local),
2) $\hat{d}_g(\mathbf{x}_i)$: isotropic Gaussian kernel with a *fixed* bandwidth calculated as the mean distance to a point's first neighbour (proposed estimator - global),
3) Isotropic Gaussian kernel with a *varying* bandwidth equal to the distance to the first nearest neighbour of each point:

$$\sigma_3(\mathbf{x}_i) = \hat{f}_{s;l} = \delta_{i;1}. \tag{10}$$

Note that the only difference between this estimator and ours is the bandwidth normalisation by π.

[1] A probability density function can be seen as a (continuous) signal and, as such, it can be decomposed in terms of frequencies.

Table I
COMPARISON OF DENSITY ESTIMATORS IN TERMS OF DI (SMALLER IS BETTER).

	Pavia U.	Indian Pines	KSC	Salinas	Bostwana	Massey
Proposed (local)	**0.53**	**1.09**	**0.56**	**0.72**	**0.59**	**0.90**
Proposed (global)	0.99	1.28	0.80	0.94	1.01	1.22
Kernel (local dist. to NN)	0.85	1.17	0.68	0.91	0.95	0.98
Kernel (rule of thumb)	1.33	1.33	1.19	1.32	1.34	1.31
Inv. dist. to NN	0.97	1.28	0.79	0.91	1.01	1.22
Diffusion	1.31	1.34	1.09	1.27	1.32	1.29
Self-consistent	1.32	1.30	0.90	1.25	1.35	1.31

Table II
COMPARISON OF DENSITY ESTIMATORS IN TERMS OF MINIMAX SILHOUETTE (BETWEEN -1 AND 1, LARGER IS BETTER - POSITIVE VALUES MEAN THAT CLUSTERS DO NOT OVERLAP ON AVERAGE).

	Pavia U.	Indian Pines	KSC	Salinas	Bostwana	Massey
Proposed (local)	0.00	**-0.35**	0.02	0.03	-0.01	-0.30
Proposed (global)	0.06	-0.36	**0.09**	0.16	0.02	**-0.26**
Kernel (local dist. to NN)	-0.01	**-0.35**	0.04	0.02	-0.02	-0.30
Kernel (rule of thumb)	-0.01	-0.36	-0.03	0.02	-0.02	-0.27
Inv. dist. to NN	**0.07**	-0.36	**0.09**	**0.17**	**0.02**	**-0.26**
Diffusion	-0.02	-0.36	-0.05	-0.02	-0.03	-0.28
Self-consistent	-0.01	**-0.35**	0.07	0.06	-0.02	-0.28

4) Isotropic Gaussian kernel with a *fixed* bandwidth determined according to Silverman's rule of thumb [23]:

$$\sigma_4 = \left[\frac{4}{D+2} \right]^{\frac{1}{D+4}} N^{\frac{1}{D+4}} s_i \Bigg]^2, \qquad (11)$$

where s_i is the standard deviation of the data in dimension i (in our case: $s_i = 1, \forall i$).

5) Inverse of the distance to first neighbour,

$$\hat{d}_5(\mathbf{x}_i) = \frac{1}{\delta_{i;1}}. \qquad (12)$$

6) Diffusion-based kernel estimator [19] (*varying* bandwidth),
7) Self-consistent kernel estimator [24] (*varying* bandwidth).

All estimators except the latter two are based on a calculation of a similarity matrix. The diffusion-based and self-consistent estimators use the Fourier transform of the data, which makes them substantially slower than the rest of the benchmark. However, they both have been reported to perform well at identifying cluster structures which is why we included them in our experiments, as a basis for comparison.

All the results reported are obtained with the squared Euclidean distance as similarity measure. We tested five different distance functions (squared Euclidean distance, regular Euclidean distance, spectral angle mapper, L-0.5 norm and Euclidean distance between the pixels' spectral derivative) and found that none of them performs systematically and significantly better than the others in terms of the criteria used in this study.

B. Data

We used 6 hyperspectral images: Pavia University, Indian Pines, Kennedy Space Center, Salinas, Bostwana and Massey University. Their characteristics are given in Table III.

Table III
DESCRIPTION OF DATA SETS USED: NUMBER OF PIXELS N, DIMENSIONALITY D AND NUMBER OF CLASSES.

Data set	N	D	Classes
Pavia University	42776	103	9
Indian Pines	10249	200	16
Kennedy Space Center	5211	176	13
Salinas	54129	224	16
Bostwana	3229	111	14
Massey University	9564	339	23

Calculating the density of each pixels in large datasets such as Pavia University and Salinas is practically intractable on a desktop computer, mostly due to the memory complexity incurred by the computation of the similarity matrix. For each data set with more than 10,000 pixels, we performed 20 random selections[2] of 10,000 points and computed the average results.

C. Results

Results of DI scores are given in Table I. In Table II, we give the average silhouette indices calculated from the ground truth data over the range of τ. Large values indicate good overall cluster separability and an efficient removal of low-density pixels. To calculate the silhouette, we used a graph-based measure of similarity known as the minimax distance [25]. This measure is more meaningful in the study of arbitrary-shaped clusters as it results in significantly larger average silhouette indices than with regular distances.

[2]Note that the same proportion of each class was kept each time.

Examples of $\Delta(\tau)$ are given in Fig. 2. Recall that $\Delta(\tau)$ graphs should be as close as possible to the diagonal.

These results clearly indicate that the proposed local density estimator $\hat{d}_l(\mathbf{x}_i)$ is substantially more adaptive to variations in cluster size and sparsity than the benchmark as it outperforms all other estimators by a significant margin in terms of DI, in all scenes. This means that it distributes its energy more evenly to all clusters, thus giving them all equal importance. It is particularly noteworthy that estimator 3 (Kernel - local dist. to NN), which also uses the distance to a pixel's first neighbour as the local kernel bandwidth (but without normalisation) performs significantly worse than our estimator in terms of DI overall. The self-consistent estimator also performs well on Pavia U., KSC and Bostwana, but at the expense of computation time. Actually, we found that both the diffusion-based and self-consistent estimator are about 4 times slower than those that use a similarity matrix, with an average of 30 seconds needed to process 10,000 points, while kernel density estimators require about 7 seconds, most of which is dedicated to the calculation of the similarity matrix.

The proposed global estimator $\hat{d}_g(\mathbf{x}_i)$, on the other hand performs slightly worse in terms of adaptability to local changes in sparsity, which was expected. However, it also surpasses the rule of thumb, diffusion, and self-consistent estimators. This is noteworthy as the latter two have an adaptive bandwidth. One possible explanation for this is that 1) the rule of thumb estimator is simply too

In terms of silhouette, generally the values are fairly low. The fact that some of them they are negative indicates that the corresponding density estimators concentrate their energy in regions where several cluster overlap. This is mostly due to the high dimensionality of the data, which makes it difficult to rely on any measure of distance to determine similarity. As a result, clusters can have very arbitrary shapes and so can distributions of distances within clusters. However, the silhouette index can still be used as a means to compare different clustering results or, in our case, density estimators. Even in high dimensionality, a higher silhouette index is desirable. What we can conclude from these results is that the proposed estimators do not perform significantly worse overall than the benchmark. The estimator with globally estimated sampling frequency $\hat{d}_g(\mathbf{x}_i)$ seems to give a better class-separability than $\hat{d}_l(\mathbf{x}_i)$ overall, while the best results are obtained when using the inverse to the nearest neighbour as a measure of density.

In conclusion, the proposed parameter-free estimators can help find representative core points more accurately than others from the benchmark, particularly $\hat{d}_l(\mathbf{x}_i)$, which is more adaptive than $\hat{d}_g(\mathbf{x}_i)$, even though it yields a slightly worse class-separability.

V. Conclusions

In this paper, we looked at the problem of density estimation for clustering hyperspectral pixels. We proposed a new criterion, the Degree of Imbalance, to estimate the adaptability of a density estimator to clusters of varied size and sparsity. We also proposed a new automatic bandwidth selection heuristic for kernel-based estimator that is based on the Nyquist frequency. We demonstrated that the proposed estimator is significantly more adaptive than most existing parameter-free estimators and that should then be considered in core points detection and density ranking-based clustering methods. Another critical aspect of core point selection is the ability to increase the separability of clusters. In other words, the smaller the set of core points, the easier the clustering should be. Results of calculating a modified silhouette index demonstrate empirically that the proposed estimators perform well at separating clusters. Future work will focus on further determining how to get a compromise between the three objectives for density estimation: adaptability, class-separability, and speed.

References

[1] K. Chehdi and C. Cariou, "The true false ground truths: What interest?," in *Image and Signal Processing for Remote Sensing XXII*. International Society for Optics and Photonics, 2016, vol. 10004, p. 100040M.

[2] O. Grygorash, Y. Zhou, and Z. Jorgensen, "Minimum spanning tree based clustering algorithms," in *Tools with Artificial Intelligence, 2006. ICTAI'06. 18th IEEE International Conference on*. IEEE, 2006, pp. 73–81.

[3] T. Kanungo, D. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 7, pp. 881–892, 2002.

[4] H.P. Kriegel, P. Kröger, J. Sander, and A. Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[5] Y. Cheng, "Mean shift, mode seeking, and clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 17, no. 8, pp. 790–799, 1995.

[6] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[7] 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 *Kdd*, 1996, vol. 96, pp. 226–231.

[8] C. Genovese, M. Perone-Pacifico, I. Verdinelli, L. Wasserman, et al., "Nonparametric ridge estimation," *The Annals of Statistics*, vol. 42, no. 4, pp. 1511–1545, 2014.

[9] M. Du, Sh. Ding, and H. Jia, "Study on density peaks clustering based on k-nearest neighbors and principal component analysis," *Knowledge-Based Systems*, vol. 99, pp. 135–145, 2016.

[10] X. Yang, Q. Zhu, Y. Huang, J. Xiao, L. Wang, and F. Tong, "Parameter-free laplacian centrality peaks clustering," *Pattern Recognition Letters*, vol. 100, pp. 167–173, 2017.

[11] J. Lu and Q. Zhu, "An effective algorithm based on density clustering framework," *IEEE Access*, vol. 5, pp. 4991–5000, 2017.

[12] C. Cariou and K. Chehdi, "Nearest neighbor-density-based clustering methods for large hyperspectral images," in *Image and Signal Processing for Remote Sensing XXIII*. International Society for Optics and Photonics, 2017, vol. 10427, p. 104270I.

[13] D. Loftsgaarden, C. Quesenberry, et al., "A nonparametric estimate of a multivariate density function," *The Annals of Mathematical Statistics*, vol. 36, no. 3, pp. 1049–1051, 1965.

[14] T. Tran, R. Wehrens, and L. Buydens, "Knn-kernel density-based clustering for high-dimensional multivariate data," *Computational Statistics & Data Analysis*, vol. 51, no. 2, pp. 513–525, 2006.

[15] J. Stevens, R. Resmini, and D. Messinger, "Spectral-density-based graph construction techniques for hyperspectral image analysis," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5966–5983, 2017.

[16] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065–1076, 1962.