Kernel clustering: density biases and solutions

Dmitrii Marin* Meng Tang* Ismail Ben Ayed† Yuri Boykov*

*Computer Science, University of Western Ontario, Canada †École de Technologie Supérieure, University of Quebec, Canada
dmitrii.a.marin@gmail.com mtang73@csd.uwo.ca ismail.benayed@etsmtl.ca yuri@csd.uwo.ca

Abstract— Clustering is widely used in data analysis where kernel methods are particularly popular due to their generality and discriminating power. However, kernel clustering has a practically significant bias to small dense clusters, *e.g.* empirically observed in [1]. Its causes have never been analyzed and understood theoretically, even though many attempts were made to improve the results. We provide conditions and formally prove this bias in kernel clustering. Previously, Breiman [2] proved a bias to histogram mode isolation in *discrete* Gini criterion for decision tree learning. We found that kernel clustering reduces to a *continuous* generalization of Gini criterion for a common class of kernels where we prove a bias to density mode isolation and call it *Breiman's bias*. These theoretical findings suggest that a principal solution for the bias should directly address data density inhomogeneity. In particular, we show that *density equalization* can be implicitly achieved using either locally adaptive weights or a general class of Riemannian (geodesic) kernels. Our density equalization principle unifies many popular kernel clustering criteria including *normalized cut*, which we show has a bias to sparse subsets inversely related to Breiman's bias. Our synthetic and real data experiments illustrate these density biases and proposed solutions. We anticipate that theoretical understanding of kernel clustering limitations and their principled solutions will be important for a broad spectrum of data analysis applications across the disciplines.

1 Introduction

In machine learning, *kernel clustering* is a well established data analysis technique [3], [4], [1], [5], [6], [7], [8], [9], [10], [11] that can identify non-linearly separable structures, see Figure 1(a-b). Section 1.1 reviews the kernel K-means and related clustering objectives, some of which have theoretically explained biases, see Section 1.2. In particular, Section 1.2.2 describes the discrete *Gini clustering criterion* standard in decision tree learning where Breiman [2] proved a bias to histogram mode isolation.

Empirically, it is well known that kernel K-means or *average association* (see Section 1.1.1) has a bias to so-called "tight" clusters for small bandwidths [1]. Figure 1(c) demonstrates this bias on a non-uniform modification of a typical toy example for kernel K-means with common Gaussian kernel

$$k(x,y) \propto \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right). \tag{1}$$

This paper shows in Section 2 that under certain conditions kernel K-means approximates the *continuous* generalization of the Gini criterion where we formally prove a mode isolation bias similar to the discrete case analyzed by Breiman. Thus, we refer to the "tight" clusters in kernel K-means as *Breiman's bias*.

We propose a *density equalization* principle directly addressing the cause of Breiman's bias. First, Section 3 discusses modification of the density with adaptive point weights. Then, Section 4 proposes a general class of locally adaptive *geodesic kernels*

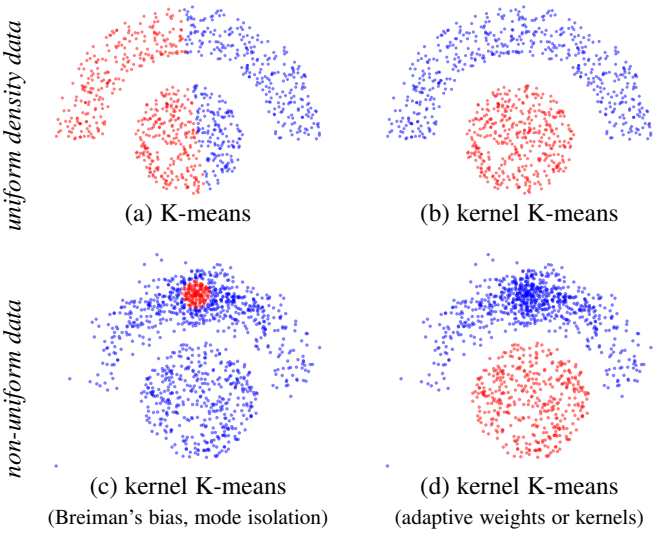


Fig. 1: Kernel K-means with Gaussian kernel (1) gives desirable nonlinear separation for *uniform* density clusters (a,b). But, for *non-uniform* clusters in (c) it either isolates a small dense "clump" for smaller σ due to Breiman's bias (Section 2) or gives results like (a) for larger σ. No fixed σ yields solution (d) given by locally adaptive kernels or weights eliminating the bias (Sections 4 & 3).

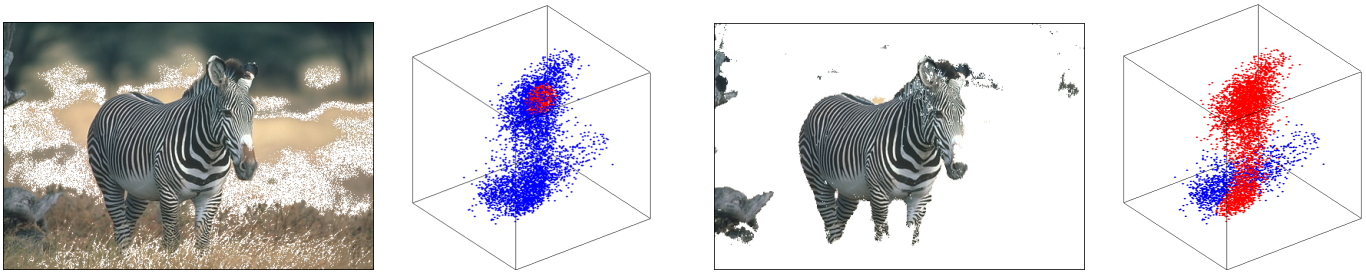
implicitly transforming data and modifying its density. We derive "density laws" relating adaptive weights and kernels to density transformations. They allow to implement *density equalization* resolving Breiman's bias, see Figure 1(d). One popular heuristic [12] approximates a special case of our Riemannian kernels.

Besides mode isolation, kernel clustering may have the opposite density bias, *e.g. sparse subsets* in Normalized Cut [1], see Figure 9(a). Section 5 presents "normalization" as implicit data density inversion establishing a formal relation between sparse subsets and Breiman's bias. Equalization addresses any density biases. Interestingly, density equalization makes many standard pairwise clustering criteria conceptually equivalent, see Section 6.

1.1 Kernel K-means

A popular data clustering technique, *kernel K-means* [3] is a generalization of the basic *K-means* method. Assuming Ω denotes a finite set of points and $f_p \in \mathcal{R}^N$ is a feature (vector) for point p, the basic K-means minimizes the sum of squared errors within clusters, that is, distances from points f_p in each cluster $S_k \subset \Omega$ to the cluster means m_k

$$\left(\begin{array}{c}\textbf{k-means}\\ \textbf{criterion}\end{array}\right) \qquad \sum_k \sum_{p \in S^k} \|f_p - m_k\|^2. \tag{2}$$



(a) Breiman's bias (b) good clustering

Fig. 2: Example of Breiman's bias on real data. Feature vectors are 3-dimensional LAB colours corresponding to image pixels. Clustering results are shown in two ways. First, *red* and *blue* show different clusters inside LAB space. Second, pixels with colours in the "background" (red) cluster are removed from the original image. (a) shows the result for kernel K-means with a fixed-width Gaussian kernel isolating a small dense group of pixels from the rest. (b) shows the result for an adaptive kernel, see Section 4.

Instead of clustering data points $\{f_p \mid p \in \Omega\} \subset \mathcal{R}^N$ in their original space, kernel K-means uses mapping $\phi : \mathcal{R}^N \to \mathcal{H}$ embedding input data $f_p \in \mathcal{R}^N$ as points $\phi_p \equiv \phi(f_p)$ in a higher-dimensional Hilbert space $\mathcal{H}$. Kernel K-means minimizes the sum of squared errors in the embedding space corresponding to the following (mixed) objective function

$$F(S,m) = \sum_k \sum_{p \in S^k} \|\phi_p - m_k\|^2 \tag{3}$$

where $S = (S^1, S^2, \ldots, S^K)$ is a partitioning (clustering) of Ω into K clusters, $m = (m_1, m_2, \ldots m_K)$ is a set of parameters for the clusters, and $\|.\|$ denotes the Hilbertian norm[1]. Kernel K-means finds clusters separated by hyperplanes in $\mathcal{H}$. In general, these hyperplanes correspond to non-linear surfaces in the original input space $\mathcal{R}^N$. In contrast to (3), standard K-means objective (2) is able to identify only linearly separable clusters in $\mathcal{R}^N$.

Optimizing F with respect to the parameters yields closed-form solutions corresponding to the cluster means in the embedding space:

$$\hat{m}_k = \frac{\sum_{q \in S^k} \phi_q}{|S^k|} \tag{4}$$

where $|.|$ denotes the cardinality (number of points) in a cluster. Plugging optimal means (4) into objective (3) yields a high-order function, which depends solely on the partition variable S:

$$F(S) = \sum_k \sum_{p \in S^k} \left\| \phi_p - \frac{\sum_{q \in S^k} \phi_q}{|S^k|} \right\|^2. \tag{5}$$

Expanding the Euclidean distances in (5), one can obtain an equivalent pairwise clustering criterion expressed solely in terms of inner products $\langle \phi(f_p), \phi(f_q) \rangle$ in the embedding space $\mathcal{H}$:

$$F(S) \stackrel{c}{=} -\sum_k \frac{\sum_{pq \in S^k} \langle \phi(f_p), \phi(f_q) \rangle}{|S^k|} \tag{6}$$

where $\stackrel{c}{=}$ means equality up to an additive constant. The inner product is often replaced with kernel k, a symmetric function:

$$k(x,y) := \langle \phi(x), \phi(y) \rangle. \tag{7}$$

Then, kernel K-means objective (5) can be presented as

$$\left(\begin{array}{c} \textbf{kernel} \\ \textbf{k-means} \\ \textbf{criterion} \end{array} \right) \qquad F(S) \stackrel{c}{=} -\sum_k \frac{\sum_{pq \in S^k} k(f_p, f_q)}{|S^k|}. \tag{8}$$

1. Our later examples use finite-dimensional embeddings ϕ where $\mathcal{H} = \mathcal{R}^M$ is an Euclidean space ($M \gg N$) and $\|.\|$ is the Euclidean norm.

Formulation (8) enables optimization in high-dimensional space $\mathcal{H}$ that only uses kernel computation and does not require computing the embedding $\phi(x)$. Given a kernel function, one can use the kernel K-means without knowing the corresponding embedding. However, not any symmetric function corresponds to the inner product in some space. Mercer's theorem [4] states that any *positive semidefinite* (p.s.d.) kernel function $k(x,y)$ can be expressed as an inner product in a higher-dimensional space. While p.s.d. is a common assumption for kernels, pairwise clustering objective (8) is often extended beyond p.s.d. affinities. There are many other extension of kernel K-means criterion (8). Despite the connection to density modes made in our paper, kernel clustering has only a weak relation to *mean-shift* [13], *e.g.* see [14].

1.1.1 Related graph clustering criteria

Positive semidefinite kernel $k(f_p, f_q)$ in (8) can be replaced by an arbitrary pairwise similarity or affinity matrix $A = [A_{pq}]$. This yields the *average association* criterion, which is known in the context of graph clustering [1], [15], [8]:

$$-\sum_k \frac{\sum_{pq \in S^k} A_{pq}}{|S^k|}. \tag{9}$$

The standard kernel K-means algorithm [8], [10] is not guaranteed to decrease (9) for improper (non p.s.d.) kernel $k := A$. However, [15] showed that dropping p.s.d. assumption is not essential: for arbitrary association A there is a p.s.d. kernel k such that objective (8) is equivalent to (9) up to a constant.

In [1] authors experimentally observed that the average association (9) or kernel K-means (8) objectives have a bias to separate small dense group of data points from the rest, *e.g.* see Figure 2.

Besides average association, there are other pairwise graph clustering criteria related to kernel K-means. *Normalized cut* is a common objective in the context of spectral clustering [1], [16]. It optimizes the following objective

$$-\sum_k \frac{\sum_{pq \in S^k} A_{pq}}{\sum_{p \in S^k} d_p}. \tag{10}$$

where $d_p = \sum_{q \in \Omega} A_{pq}$. Note that for $d_p = 1$ equation (10) reduces to (9). It is known that Normalized cut objective is equivalent to a weighted version of kernel K-means criterion [17], [8].

1.1.2 Probabilistic interpretation via kernel densities

Besides *kernel clustering*, kernels are also commonly used for *probability density estimation*. This section relates these two independent problems. Standard *multivariate kernel density estimate*

or *Parzen density estimate* for the distribution of data points within cluster S^k can be expressed as follows [18]:

$$\mathcal{P}_\Sigma(x|S^k) \;:=\; \frac{\sum_{q\in S^k} k(x,f_q)}{|S^k|}, \tag{11}$$

with kernel k having the form:

$$k(x,y) \;=\; |\Sigma|^{-\frac{1}{2}}\,\psi\left(\Sigma^{-\frac{1}{2}}(x-y)\right) \tag{12}$$

where ψ is a symmetric multivariate density and Σ is a symmetric positive definite *bandwidth* matrix controlling the density estimator's smoothness. One standard example is the Gaussian (normal) kernel (1) corresponding to

$$\psi(t) \;\propto\; \exp\left(-\frac{\|t\|^2}{2}\right), \tag{13}$$

which is commonly used both in kernel density estimation [18] and kernel clustering [7], [1].

The choice of bandwidth Σ is crucial for accurate density estimation, while the choice of l plays only a minor role [19]. There are numerous works regarding optimal bandwidth selection [20], [19], [21]. For example, Scott's *rule of thumb* is

$$\sqrt{\Sigma_{ii}} = \frac{r_i}{\sqrt[N+4]{n}}, \qquad \Sigma_{ij} = 0 \text{ for } i \neq j \tag{14}$$

where n is the number of points, and r_i^2 is the variance of the i-th feature that could be interpreted as the range or scale of the data. Scott's rule gives optimal *mean integrated squared error* for normal data distribution, but in practice it works well in more general settings. In all cases the optimal bandwidth for sufficiently large datasets is a small fraction of the data range [22], [18]. For shortness, we use adjective *r-small* to describe bandwidths providing accurate density estimation.

If kernel k has form (12) up to a positive multiplicative constant then kernel K-means objective (8) can be expressed in terms of kernel densities (11) for points in each cluster [7]:

$$F(S) \;\overset{c}{=}\; -\sum_k \sum_{p\in S^k} \mathcal{P}_\Sigma(f_p|S^k). \tag{15}$$

1.2 Other clustering criteria and their known biases

One of the goals of this paper is a theoretical explanation for the bias of kernel K-means with small bandwidths toward tight dense clusters, which we call *Breiman's bias*, see Figs 1-2. This bias was observed in the past only empirically. As discussed in Section 4.1, large bandwidth reduces kernel K-means to basic K-means where bias to equal cardinality clusters is known [23]. This section reviews other standard clustering objectives, entropy and Gini criteria, that have biases already well-understood theoretically. In Section 2 we establish a connection between Gini clustering and kernel K-means in case of *r-small* kernels. This connection allows theoretical analysis of Breiman's bias in kernel K-means.

1.2.1 Probabilistic K-means and entropy criterion

Besides non-parametric kernel K-means clustering there are well-known parametric extensions of basic K-means (2) based on probability models. *Probabilistic K-means* [23] or *model based clustering* [24] use some given likelihood functions $P(f_p|\theta_k)$ instead of distances $\|f_p-\theta_k\|^2$ in (2) as in clustering objective

$$-\sum_k \sum_{p\in S^k} \log P(f_p|\theta_k). \tag{16}$$

Note that objective (16) reduces to basic K-means (2) for Gaussian probability model $P(.|\theta_k)$ with mean θ_k and a fixed scalar covariance matrix.

In probabilistic K-means (16) models can differ from Gaussians depending on *a priori* assumptions about the data in each cluster, *e.g.* gamma, Gibbs, or other distributions can be used. For more complex data, each cluster can be described by highly-descriptive parametric models such as Gaussian mixtures (GMM). Instead of kernel density estimates in kernel K-means (15), probabilistic K-means (16) uses parametric distribution models. Another difference is the absence of the log in (15) compared to (16).

The analysis in [23] shows that in case of highly descriptive model P, *e.g.* GMM or histograms, (16) can be approximated by the standard *entropy criterion* for clustering:

$$\left(\begin{array}{c}\textbf{entropy}\\ \textbf{criterion}\end{array}\right) \qquad \sum_k |S^k|\cdot H(S^k) \tag{17}$$

where $H(S^k)$ is the entropy of the distribution of the data in S^k:

$$H(S^k) \;:=\; -\int P(x|\theta_k)\log P(x|\theta_k)\,\mathrm{d}x.$$

The discrete version of the entropy criterion is widely used for learning binary decision trees in classification [2], [18], [25]. It is known that the entropy criterion above is biased toward equal size clusters [2], [23], [26].

1.2.2 Discrete Gini impurity and criterion

Both Gini and entropy clustering criteria are widely used in the context of decision trees [18], [25]. These criteria are used to decide the best split at a given node of a binary classification tree [27]. The Gini criterion can be written for clustering $\{S^k\}$ as

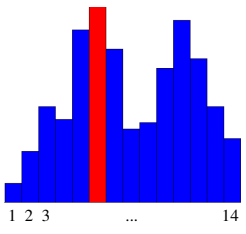
$$\left(\begin{array}{c}\textbf{discrete}\\ \textbf{Gini criterion}\end{array}\right) \qquad \sum_k |S^k|\cdot G(S^k) \tag{18}$$

where $G(S^k)$ is the *Gini impurity* for the points in S^k. Assuming discrete feature space $\mathcal{L}$ instead of $\mathcal{R}^N$, the Gini impurity is

$$G(S^k) \;:=\; 1-\sum_{l\in\mathcal{L}} \mathcal{P}(l\,|S^k)^2 \tag{19}$$

where $\mathcal{P}(\cdot|S^k)$ is the empirical probability (histogram) of discrete-valued features $f_p \in \mathcal{L}$ in cluster S^k.

Similarly to the entropy, Gini impurity $G(S^k)$ can be viewed as a measure of sparsity or "peakedness" of the distribution for points in S^k. Note that (18) has a form similar to the entropy criterion in (17), except that entropy H is replaced by the Gini impurity. Breiman [2] analyzed the theoretical properties of the discrete Gini criterion (18) when $\mathcal{P}(\cdot|S^k)$ are *discrete histograms*. He proved



Theorem 1 (Breiman)**.** *For $K=2$ the minimum of the Gini criterion (18) for discrete Gini impurity (19) is achieved by assigning all data points with the highest-probability feature value in $\mathcal{L}$ to one cluster and the remaining data points to the other cluster, as in example for $\mathcal{L}=\{1,\ldots,14\}$ on the left.* □

2 BREIMAN'S BIAS (NUMERICAL FEATURES)

In this section we show that the kernel K-means objective reduces to *continuous* Gini criterion under some general conditions on the kernel function. We further formally prove that the optimum of the continuous Gini criterion separates a small group of data points from the rest. That is, we show that the discussed earlier biases observed in the context of clustering [1] and decision tree learning [2] are the same phenomena.

For further analysis we reformulate the problem of clustering a discrete set of points $\{f_p \,|\, p \in \Omega\} \subset \mathcal{R}^N$, see Section 1.1, as a continuous domain clustering problem. Let ρ be a continuous probability density function over domain $\mathcal{R}^N$ such that the discrete points f_p could be treated as samples from this distribution. The clustering of the continuous domain will be described by an *assignment function* $s : \mathcal{R}^N \to \{1, 2, \ldots, K\}$. Density ρ implies conditional probability densities $\rho_k^s(x) := \rho(x \,|\, s(x) = k)$. Feature points f_p in cluster S^k could be interpreted as a sample from conditional density ρ_k^s.

Then, the continues clustering problem is to find an assignment function optimizing some clustering criteria. For example, we can analogously to (18) define continuous Gini clustering criterion

$$\left(\begin{matrix}\textbf{continuous}\\ \textbf{Gini criterion}\end{matrix}\right) \qquad \sum_k w_k \cdot G(s,k), \tag{20}$$

where w_k is the probability to draw a point from k-th cluster and

$$G(s,k) \;:=\; 1 - \int \rho_k^s(x)^2 \,\mathrm{d}x. \tag{21}$$

In the next section we show that kernel K-means energy (15) can be approximated by continuous Gini-clustering criterion (20) for *r-small* kernels.

2.1 Kernel K-means and continuous Gini criterion

To establish the connection between kernel clustering and the Gini criterion, let us first recall Monte-Carlo estimation [23], which yields the following expectation-based approximation for a continuous function $g(x)$ and cluster $C \subset \Omega$:

$$\sum_{p \in C} g(f_p) \approx |C| \int g(x)\, \rho_C(x)\; \mathrm{d}x \tag{22}$$

where ρ_C is the "true" continuous density of features in cluster C. Using (22) for $C = S^k$ and $g(x) = \mathcal{P}_\Sigma(x|S^k)$, we can approximate the kernel density formulation in (15) by its expectation

$$F(S) \;\overset{c}{\approx}\; -\sum_k |S^k| \int \mathcal{P}_\Sigma(x|S^k)\, \rho_k^s(x)\; \mathrm{d}x. \tag{23}$$

Note that partition $S = (S^1, \ldots, S^K)$ is determined by dataset Ω and assignment function s. If kernel k has *r-small* bandwidth Σ optimal for accurate kernel density estimation, we have

$$\mathcal{P}_\Sigma(\cdot\,|S^k) \;\approx\; \rho_k^s(\cdot) \tag{24}$$

further reducing (23) to an approximation

$$F(S) \;\overset{c}{\approx}\; -\sum_k |S^k| \cdot \int \rho_k^s(x)^2 \,\mathrm{d}x \;\overset{c}{\equiv}\; \sum_k |S^k| \cdot G(s,k). \tag{25}$$

Additional application of Monte-Carlo estimation $|S^k|/|\Omega| \approx w_k$ allows replacing set cardinality $|S^k|$ by probability w_k of drawing a point from S^k. This results in continuous Gini clustering criterion (20), which approximates (15) or (8) up to an additive and positive multiplicative constants.

Next section proves that the continuous Gini criterion (20) has a similar bias observed by Breiman in the discrete case.

2.2 Breiman's bias in continuous Gini criterion

This section extends Theorem 1 to continuous Gini criterion (20). Section 2.1 has already established a close relation between continuous Gini criterion and kernel K-means with *r-small* bandwidth kernels. Thus, Breiman's bias also applies to the latter.

Theorem 2 (Breiman's bias in continuous case)**.** *For $K = 2$ the continuous Gini clustering criterion (20) achieves its optimal value at the partitioning of $\mathcal{R}^N$ into regions*

$$s_1 = \arg\max_x \rho(x) \quad \textit{and} \quad s_2 = \mathcal{R}^N \setminus s_1.$$

Proof. The statement follows from Lemma 2 below. □

We denote mathematical expectation of function $z : \Omega \to \mathcal{R}^1$

$$\mathbf{E}z \;:=\; \int z(x)\rho(x)\,\mathrm{d}x.$$

Minimization of (20) corresponds to maximization of the following objective function

$$L(s) \;:=\; w \int \rho_1^s(x)^2\,\mathrm{d}x \;+\; (1-w)\int \rho_2^s(x)^2\,\mathrm{d}x \tag{26}$$

where

$$w \;:=\; w_1 \;=\; \int_{s(x)=1} \rho(x)\,\mathrm{d}x = \mathbf{E}[s(x)=1]$$

where $[\cdot]$ is the indicator function. Note that conditional density ρ_1^s in (26) can be written as

$$\rho_1^s(x) \;=\; \rho(x) \cdot \frac{[s(x)=1]}{w}. \tag{27}$$

Equations (26) and (27) give

$$\begin{aligned} L(s) \;&=\; \frac{1}{w}\int \rho(x)^2 [s(x)=1]\,\mathrm{d}x \\ &+\; \frac{1}{1-w}\int \rho(x)^2 [s(x)=2]\,\mathrm{d}x. \end{aligned} \tag{28}$$

Introducing notation

$$I \;:=\; [s(x)=1] \qquad \text{and} \qquad \rho \;:=\; \rho(x)$$

allows to further rewrite objective function $L(s)$ as

$$L(s) \;=\; \frac{\mathbf{E}I\rho}{\mathbf{E}I} \;+\; \frac{\mathbf{E}(1-I)\rho}{1-\mathbf{E}I}. \tag{29}$$

Without loss of generality assume that $\frac{\mathbf{E}(1-I)\rho}{1-\mathbf{E}I} \le \frac{\mathbf{E}I\rho}{\mathbf{E}I}$ (the opposite case would yield a similar result). We now need following

Lemma 1. *Let a, b, c, d be some positive numbers, then*

$$\frac{a}{b} \le \frac{c}{d} \;\Longrightarrow\; \frac{a}{b} \le \frac{a+c}{b+d} \le \frac{c}{d}.$$

Proof. Use reduction to a common denominator. □

Lemma 1 implies inequality

$$\frac{\mathbf{E}(1-I)\rho}{1-\mathbf{E}I} \;\le\; \mathbf{E}\rho \;\le\; \frac{\mathbf{E}I\rho}{\mathbf{E}I}, \tag{30}$$

which is needed to prove the Lemma below.

Lemma 2. *Assume that function $s_\varepsilon : \mathcal{D} \to \{1, 2\}$ is*

$$s_\varepsilon(x) \;:=\; \begin{cases} 1, & \rho(x) \ge \sup_x \rho(x) - \varepsilon, \\ 2, & \textit{otherwise}. \end{cases} \tag{31}$$

Then

$$\sup_s L(s) = \lim_{\varepsilon \to 0} L(s_\varepsilon) = \mathbf{E}\rho + \sup_x \rho(x). \tag{32}$$

Proof. Due to monotonicity of expectation we have

$$\frac{\mathbf{E}FI}{\mathbf{E}I} \leq \frac{\mathbf{E}\left(I \sup_x \rho(x)\right)}{\mathbf{E}I} = \sup_x \rho(x). \tag{33}$$

Then (30) and (33) imply

$$L(s) = \frac{\mathbf{E}I\rho}{\mathbf{E}I} + \frac{\mathbf{E}(1-I)\rho}{1-\mathbf{E}I} \leq \sup_x \rho(x) + \mathbf{E}\rho. \tag{34}$$

That is, the right part of (32) is an upper bound for $L(s)$.

Let $I_\varepsilon \equiv [s_\varepsilon(x) = 1]$. It is easy to check that

$$\lim_{\varepsilon\to 0} \frac{\mathbf{E}(1-I_\varepsilon)\rho}{1-\mathbf{E}I_\varepsilon} = \mathbf{E}\rho. \tag{35}$$

Definition (31) also implies

$$\lim_{\varepsilon\to 0} \frac{\mathbf{E}I_\varepsilon\rho}{\mathbf{E}I_\varepsilon} \geq \lim_{\varepsilon\to 0} \frac{\mathbf{E}(\sup_x \rho(x) - \varepsilon)I_\varepsilon}{\mathbf{E}I_\varepsilon} = \sup_x \rho(x). \tag{36}$$

This result and (33) conclude that

$$\lim_{\varepsilon\to 0} \frac{\mathbf{E}I_\varepsilon\rho}{\mathbf{E}I_\varepsilon} = \sup_x \rho(x). \tag{37}$$

Finally, the limits in (35) and (37) imply

$$\begin{aligned}\lim_{\varepsilon\to 0} L(s_\varepsilon) &= \lim_{\varepsilon\to 0} \frac{\mathbf{E}(1-I_\varepsilon)\rho}{1-\mathbf{E}I_\varepsilon} + \lim_{\varepsilon\to 0} \frac{\mathbf{E}I_\varepsilon\rho}{\mathbf{E}I_\varepsilon} \\ &= \mathbf{E}\rho + \sup_x \rho(x).\end{aligned} \tag{38}$$

This equality and bound (34) prove (32). □

This result states that the optimal assignment function separates the mode of the density function from the rest of the data. The proof considers case $K = 2$ for continuous Gini criterion approximating kernel K-means for *r-small* kernels. The multi-cluster version should be analogous. Empirically, $K > 2$ gives similar bias, see Figure 3. These are asymptotic results that are valid when the sizes of clusters are sufficiently large. In practice we observe that shrinking of the clusters invalidates approximation (23) preventing their complete collapse. As a result Breiman's bias yields small clusters separating the densest parts of the data from the rest, see Figures 2, 3, 7(a-d), 8.

Interestingly, there is also a relation between *maximum cliques* and *density modes*. Assume 0-1 kernel $[\|x - y\| \leq \sigma]$ with bandwidth σ. Then, kernel matrix A is a connectivity matrix corresponding to a *σ-disk graph*. Intuitively, the maximum clique on this graph should be inside a disk with the largest number of points in it, which corresponds to the density mode.

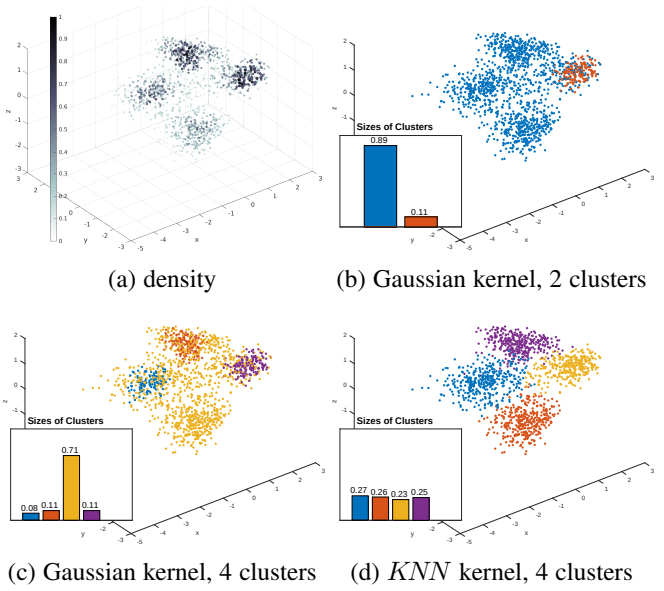
Formally, mode isolation bias can be linked to both maximum clique and its weighted-graph generalization, *dominant set* [9]. It is known that maximum clique [28] and *dominant set* [9] solve a two-region clustering problem with energy

$$-\frac{\sum_{pq\in S^1} A_{pq}}{|S^1|} \tag{39}$$

corresponding to average association (9) for $K = 1$ and $S^1 \subseteq \Omega$. Under the same assumptions as above, Gini impurity (21) can be used as an approximation reducing objective (39) to

$$\frac{\mathbf{E}I\rho}{\mathbf{E}I}. \tag{40}$$

Using (33) and (37) we can conclude that the optimum of (40) isolates the mode of density function ρ. Thus, clustering minimizing (39) for *r-small* bandwidths also has Breiman's bias. That is, for such bandwidths the concepts of maximum clique and dominant set for graphs correspond to the concept of *mode isolation* for data densities. Dominant sets for the examples in Figures 1(c), 2(a), and 7(d) would be similar to the shown mode-isolating solutions.



(a) density (b) Gaussian kernel, 2 clusters

(c) Gaussian kernel, 4 clusters (d) KNN kernel, 4 clusters

Fig. 3: Breiman's bias in clustering of images. We select 4 categories from the LabelMe dataset [29]. The last fully connected layer of the neural network in [30] gives 4096-dimensional feature vector for each image. We reduce the dimension to 5 via PCA. For visualization purposes, we obtain 3D embeddings via MDS [31]. (a) Kernel densities estimates for data points are color-coded: darker points correspond to higher density. (b,c) The result of the kernel K-means with the Gaussian kernel (1). Scott's rule of thumb defines the bandwidth. Breiman's bias causes poor clustering, *i.e.* small cluster is formed in the densest part of the data in (b), three clusters occupy few points within densest regions while the fourth cluster contains 71% of the data in (c). The *normalized mutual information* (NMI) in (c) is 0.38. (d) Good clustering produced by KNN kernel u_p (Example 3) gives NMI of 0.90, which is slightly better than the basic K-means (0.89).

3 Adaptive Weights Solving Breiman's Bias

We can use a simple modification of average association by introducing weights $w_p \geq 0$ for each point "error" within the equivalent kernel K-means objective (3)

$$F_w(S, m) = \sum_k \sum_{p\in S^k} w_p \|\phi_p - m_k\|^2. \tag{41}$$

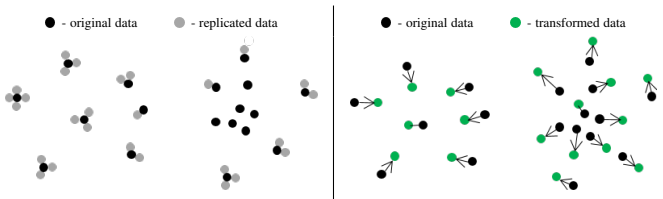
Such weighting is common for K-means [22]. Similarly to Section 1.1 we can expand the Euclidean distances in (41) to obtain an equivalent *weighted average association* criterion generalizing (9)

$$-\sum_k \frac{\sum_{pq\in S_k} w_p w_q A_{pq}}{\sum_{p\in S_k} w_p}. \tag{42}$$

Weights w_p have an obvious interpretation based on (41); they change the data by replicating each point p by a number of points in the same location (Figure 4a) in proportion to w_p. Therefore, this weighted formulation directly modifies the data density as

$$\rho'_p \propto w_p \rho_p \tag{43}$$

where ρ_p and ρ'_p are respectively the densities of the original and the new (replicated) points. The choice of $w_p = 1/\rho_p$ is a simple way for equalizing data density to solve Breiman's bias. As shown in Figure 4(a), such a choice enables low-density points to be replicated more frequently than high-density ones. This is one of density equalization approaches giving the solution in Figure 1(d).



(a) adaptive weights (Sec. 3) (b) adaptive kernels (Sec. 4.3)

Fig. 4: *Density equalization* via (a) adaptive weights and (b) adaptive kernels. In (a) the density is modified as in (43) via "replicating" each data point inverse-proportionately to the observed density using $w_p \propto 1/\rho_p$. For simplicity (a) assumes positive integer weights w_p. In (b) the density is modified according to (58) for bandwidth (61) via implicit embedding of data points in a higher dimensional space that changes their relative positions.

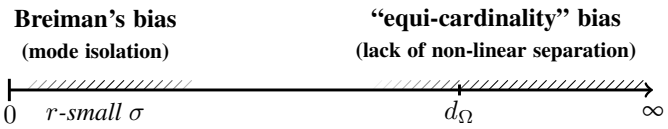


Fig. 5: Kernel K-means biases over the range of bandwidth σ. Data diameter is denoted by $d_\Omega = \max_{pq\in\Omega} \|f_p - f_q\|$. Breiman's bias is established for *r-small* σ (Section 1.1.2). Points stop interacting for σ smaller than *r-small* making kernel K-means fail. Larger σ reduce kernel K-means to the basic K-means removing an ability to separate the clusters non-linearly. In practice, there could be no intermediate good σ. In the example of Fig.1(c) any fixed σ leads to either Breiman's bias or to the lack of non-linear separability.

4 Adaptive kernels solving Breiman's bias

Breiman's bias in kernel K-means is specific to *r-small* bandwidths. Thus, it has direct implications for the bandwidth selection problem discussed in this section. Note that kernel bandwidth selection for *clustering* should not be confused with kernel bandwidth selection for *density estimation*, an entirely different problem outlined in Section 1.1.2. In fact, *r-small* bandwidths give accurate density estimation, but yield poor clustering due to Breiman's bias. Larger bandwidths can avoid this bias in clustering. However, Section 4.1 shows that for extremely large bandwidths kernel K-means reduces to standard K-means, which looses ability of non-linear cluster separation and has a different bias to equal cardinality clusters [23], [26].

In practice, avoiding extreme bandwidths is problematic since the notions of *small* and *large* strongly depend on data properties that may significantly vary across the domain. This motivates *locally* adaptive strategies. Interestingly, Section 4.2 shows that any locally adaptive bandwidth strategy implicitly corresponds to some data embedding $\Omega \to \mathcal{R}^{N'}$ deforming density of the points. That is, locally adaptive selection of bandwidth is equivalent to selection of density transformation. Local kernel bandwidth and transformed density are related via the *density law* established in (59). On the other hand, Theorem 2 and Figure 1 imply that Breiman's bias is caused by high non-uniformity of the clusters and density equalizing transformations should address it. Section 4.3 proposes adaptive kernel strategies based on our *density law* and motivated by a *density equalization* principle addressing Breiman's bias. In fact, a popular locally adaptive kernel in [12] is a special case of our density equalization principle.

4.1 Overview of extreme bandwidth cases

Section 2.1 and Theorem 2 prove that for *r-small* bandwidths the kernel K-means is biased toward "tight" clusters, as illustrated in Figures 1, 2 and 7(d). As bandwidth increases, continuous kernel density (11) no longer approximates the true distribution ρ_k^s violating (24). Thus, Gini criterion (25) is no longer valid as an approximation for kernel K-means objective (15). In practice, Breiman's bias disappears gradually as bandwidth gets larger. This is also consistent with experimental comparison of smaller and larger bandwidths in [1].

The other extreme case of bandwidth for kernel K-means comes from its reduction to basic K-means for large kernels. For simplicity, assume Gaussian kernels (1) of large bandwidth σ approaching data diameter. Then the kernel can be approximated by its Taylor expansion $\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right) \approx 1 - \frac{\|x-y\|^2}{2\sigma^2}$ and kernel K-means objective (8) for $\sigma \gg \|x-y\|$ becomes[2] (up to a constant)

$$\sum_k \frac{\sum_{pq\in S^k} \|f_p - f_q\|^2}{2\sigma^2|S^k|} \stackrel{c}{=} \frac{1}{\sigma^2} \sum_k \sum_{p\in S^k} \|f_p - m_k\|^2, \qquad (44)$$

which is equivalent to basic K-means (2) for any fixed σ.

Figure 5 summarizes kernel K-means biases for different bandwidths. For large bandwidths the kernel K-means loses its ability to find non-linear cluster separation due to reduction to the basic K-means. Moreover, it inherits the bias to equal cardinality clusters, which is well-known for the basic K-means [23], [26]. On the other hand, for small bandwidths kernel K-means has Breiman's bias proven in Section 2. To avoid the biases in Figure 5, kernel K-means should use a bandwidth neither too small nor too large. This motivates locally adaptive bandwidths.

4.2 Adaptive kernels as density transformation

This section shows that kernel clustering (8) with any *locally adaptive bandwidth* strategy satisfying some reasonable assumptions is equivalent to *fixed bandwidth* kernel clustering in a new feature space (Theorem 3) with a deformed point density. The adaptive bandwidths relate to density transformations via *density law* (59). To derive it, we interpret *adaptiveness* as non-uniform variation of distances across the feature space. In particular, we use a general concept of *geodesic kernel* defining adaptiveness via a metric tensor and illustrate it by simple practical examples.

Our analysis of Breiman's bias in Section 2 applies to general kernels (12) suitable for density estimation. Here we focus on clustering with kernels based on *radial basis functions* ψ s.t.

$$\psi(x-y) = \psi(\|x-y\|). \qquad (45)$$

To obtain adaptive kernels, we replace Euclidean metric with Riemannian inside (45). In particular, $\|x-y\|$ is replaced with *geodesic distances* $d_g(x,y)$ between features $x, y \in \mathcal{R}^N$ based on any given metric tensor $g(f)$ for $f \in \mathcal{R}^N$. This allows to define a *geodesic* or *Riemannian* kernel at any points f_p and f_q as in [11]

$$k_g(f_p, f_q) \;:=\; \psi(d_g(f_p, f_q)) \;\equiv\; \psi(d_{pq}) \qquad (46)$$

where $d_{pq} := d_g(f_p, f_q)$ is introduced for shortness.

2. Relation (44) easily follows by substituting $m_k \equiv \frac{1}{|S^k|}\sum_{p\in S^k} f_p$.

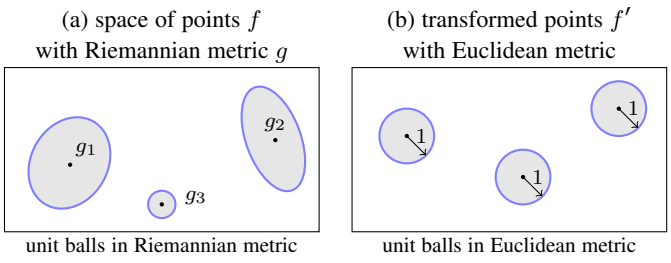


Fig. 6: Adaptive kernel (46) based on Riemannian distances (a) is equivalent to fixed bandwidth kernel after some *quasi-isometric* (50) embedding into Euclidean space (b), see Theorem 3, mapping ellipsoids (52) to balls (54) and modifying data density as in (57).

In practice, the metric tensor can be defined only at the data points $g_p := g(f_p)$ for $p \in \Omega$. Often, quickly decaying radial basis functions ψ allow Mahalanobis distance approximation inside (46)

$$d_g(f_p, x)^2 \;\approx\; (f_p - x)^T g_p (f_p - x), \tag{47}$$

which is normally valid only in a small neighborhood of f_p. If necessary, one can use more accurate approximations for $d_g(f_p, f_q)$ based on Dijkstra [32] or Fast Marching method [33].

EXAMPLE 1 (***Adaptive non-normalized***[3] ***Gaussian kernel***). *Mahalanobis distances based on (adaptive) bandwidth matrices Σ_p defined at each point p can be used to define adaptive kernel*

$$\kappa_p(f_p, f_q) \;:=\; \exp \frac{-(f_p - f_q)^T \Sigma_p^{-1} (f_p - f_q)}{2}, \tag{48}$$

which equals fixed bandwidth Gaussian kernel (1) for $\Sigma_p = \sigma^2 I$. Kernel (48) approximates (46) for exponential function ψ in (13) and tensor g continuously extending matrices Σ_p^{-1} over the whole feature space so that $g_p = \Sigma_p^{-1}$ for $p \in \Omega$. Indeed, assuming matrices Σ_p^{-1} and tensor g change slowly between points within bandwidth neighbourhoods, one can use (47) for all points in

$$\kappa_p(f_p, f_q) \;\approx\; \exp \frac{-d_g(f_p, f_q)^2}{2} \;\equiv\; \exp \frac{-d_{pq}^2}{2} \tag{49}$$

due to exponential decay outside the bandwidth neighbourhoods.

EXAMPLE 2 (***Zelnik-Manor & Perona kernel [12]***). *This popular kernel is defined as $\kappa_{pq} := \exp \frac{-\|f_p - f_q\|^2}{2\sigma_p \sigma_q}$. This kernel's relation to (46) is less intuitive due to the lack of "local" Riemannian tensor. However, under assumptions similar to those in (49), it can still be seen as an approximation of geodesic kernel (46) for some tensor g such that $g_p = \sigma_p^{-2} I$ for $p \in \Omega$. They use heuristic $\sigma_p = R_p^K$, which is the distance to the K-th nearest neighbour of f_p.*

EXAMPLE 3 (***KNN kernel***). *This adaptive kernel is defined as $u_p(f_p, f_q) = [f_q \in KNN(f_p)]$ where $KNN(f_p)$ is the set of K nearest neighbors of f_p. This kernel approximates (46) for uniform function $\psi(t) = [t < 1]$ and tensor g such that $g_p = I/(R_p^K)^2$.*

Theorem 3. *Clustering (8) with (adaptive) geodesic kernel (46) is equivalent to clustering with fixed bandwidth kernel $k'(f'_p, f'_q) := \psi'(\|f'_p - f'_q\|)$ in new feature space $\mathcal{R}^{N'}$ for some radial basis function ψ' using the Euclidean distance and some constant N'.*

3. Normalization as in (12) leads to energy (15) linked to Breiman's bias.

Proof. A powerful general result in [34], [35], [15] states that for any symmetric matrix (d_{pq}) with zeros on the diagonal there is a constant h such that squared distances

$$\widetilde{d}_{pq}^2 \;=\; d_{pq}^2 + h^2 [p \neq q] \tag{50}$$

form *Euclidean matrix* $(\widetilde{d}_{pq})$. That is, there exists some Euclidean embedding $\Omega \to \mathcal{R}^{N'}$ where for $\forall p \in \Omega$ there corresponds a point $f'_p \in \mathcal{R}^{N'}$ such that $\|f'_p - f'_q\| = \widetilde{d}_{pq}$, see Figure 6. Therefore,

$$\psi(d_{pq}) \;=\; \psi\left(\sqrt{\widetilde{d}_{pq}^2 - h^2\,[d_{pq} \geq h]}\right) \;\equiv\; \psi'(\widetilde{d}_{pq}) \tag{51}$$

for $\psi'(t) := \psi(\sqrt{t^2 - h^2[t \geq h]})$ and $k_g(f_p, f_q) = k'(f'_p, f'_q)$. □

Theorem 3 proves that *adaptive* kernels for $\{f_p\} \subset \mathcal{R}^N$ can be equivalently replaced by a *fixed* bandwidth kernel for some implicit embedding[4] $\{f'_p\} \subset \mathcal{R}^{N'}$ in a new space. Below we establish a relation between three local properties at point p: adaptive bandwidth represented by matrix g_p and two densities ρ_p and ρ'_p in the original and the new feature spaces. For $\varepsilon > 0$ consider an ellipsoid in the original space $\mathcal{R}^N$, see Figure 6(a),

$$B_p \;:=\; \{x \mid (x - f_p)^T g_p (x - f_p) \leq \varepsilon^2\}. \tag{52}$$

Assuming ε is small enough so that approximation (47) holds, ellipsoid (52) covers features $\{f_q \mid q \in \Omega_p\}$ for subset of points

$$\Omega_p \;:=\; \{q \in \Omega \mid d_{pq} \leq \varepsilon\}. \tag{53}$$

Similarly, consider a ball in the new space $\mathcal{R}^{N'}$, see Figure 6(b),

$$B'_p \;:=\; \{x \mid \|x - f'_p\|^2 \leq \varepsilon^2 + h^2\} \tag{54}$$

covering features $\{f'_q \mid q \in \Omega'_p\}$ for points

$$\Omega'_p \;:=\; \{q \in \Omega \mid \widetilde{d}_{pq}^2 \leq \varepsilon^2 + h^2\}. \tag{55}$$

It is easy to see that (50) implies $\Omega_p = \Omega'_p$. Let ρ_p and ρ'_p be the densities[5] of points within B_p and B'_p correspondingly. Assuming $|\cdot|$ denotes volumes or cardinalities of sets, we have

$$\rho_p \cdot |B_p| \;=\; |\Omega_p| \;=\; |\Omega'_p| \;=\; \rho'_p \cdot |B'_p|. \tag{56}$$

Omitting a constant factor depending on ε, h, N and N' we get

$$\rho'_p \;=\; \rho_p \frac{|B_p|}{|B'_p|} \;\propto\; \rho_p \,|\det g_p|^{-\frac{1}{2}} \tag{57}$$

representing the general form of the *density law*. For the basic isotropic metric tensor such that $g_p = I/\sigma_p^2$ it simplifies to

$$\rho'_p \;\propto\; \rho_p \,\sigma_p^N. \tag{58}$$

Thus, bandwidth σ_p can be selected adaptively based on any desired transformation of density $\rho'_p \equiv \tau(\rho_p)$ using

$$\sigma_p \;\propto\; \sqrt[N]{\tau(\rho_p)/\rho_p}. \tag{59}$$

where observed density ρ_p in the original feature space can be evaluated at any point p using any standard estimators, *e.g.* (11).

4.3 Density equalizing locally adaptive kernels

Bandwidth formula (59) works for any density transform τ. To address Breiman's bias, one can use density equalizing transforms $\tau(\rho) = const$ or $\tau(\rho) = \frac{1}{\alpha}\log(1 + \alpha\rho)$, which even up

4. The implicit embedding implied by Euclidean matrix (50) should not be confused with embedding in the Mercer's theorem for kernel methods.

5. We use the physical rather than probability density. They differ by a factor.

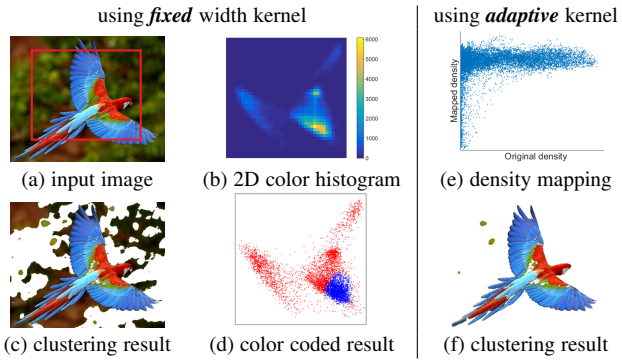
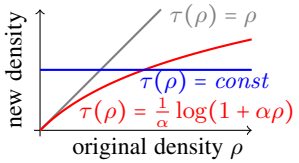


Fig. 7: (a)-(d): *Breiman's bias* for fixed bandwidth kernel (1). (f): result for (48) with adaptive bandwidth (61) s.t. $\tau(\rho) = const$. (e) *density equalization*: scatter plot of empirical densities in the original/new feature spaces obtained via (11),(50), see Appendix.

the highly dense parts of the feature space as illustrated on the right. Some empirical results using density equalization $\tau(\rho) = const$ for synthetic and real data are shown in Figures 1(d) and 7(e,f).



One way to estimate the density in (59) is KNN approach [18]

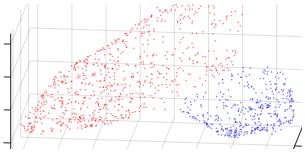
$$\rho_p \approx \frac{K}{nV_K} \propto \frac{K}{n(R_p^K)^N} \tag{60}$$

where $n \equiv |\Omega|$ is the size of the dataset, R_p^K is the distance to the K-th nearest neighbor of f_p, V_K is the volume of a ball of radius R_p^K centered at f_p. Then, density law (59) for $\tau(\rho) = const$ gives

$$\sigma_p \propto R_p^K \tag{61}$$

consistent with heuristic bandwidth in [12], see Example 2.

The result in Figure 1(d) uses adaptive Gaussian kernel (48) for $\Sigma_p = \sigma_p I$ with σ_p derived in (61). Theorem 3 claims equivalence to a fixed bandwidth kernel in some transformed higher-dimensional space $\mathcal{R}^{N'}$. Bandwidths (61) are chosen specifically to equalize the data density in this space so that $\tau(\rho) = const$. The picture on the right illustrates such density equalization for the data in Figure 1(d). It shows a 3D projection of the transformed data obtained by *multi-dimensional scaling* [31] for matrix $(\widetilde{d}_{pq})$ in (50). The observed density equalization removes Breiman's bias from the clustering in Figure 1(d).



Real data experiments for kernels with adaptive bandwidth (61) are reported in Figures 2, 3, 7, 8 and Table 1. Figure 7(e) illustrates the empirical *density equalization* effect for this bandwidth. Such data homogenization removes the conditions leading to Breiman's bias, see Theorem 2. Also, we observe empirically that KNN kernel is competitive with adaptive Gaussian kernels, but its sparsity gives efficiency and simplicity of implementation.

5 Normalized Cut and Breiman's Bias

Breiman's bias for kernel K-means criterion (8), a.k.a. *average association* (AA) (9), was empirically identified in [1], but our Theorem 2 is its first theoretical explanation. This bias was the main critique against AA in [1]. They also criticize *graph cut* [39] that "favors cutting small sets of isolated nodes". These critiques

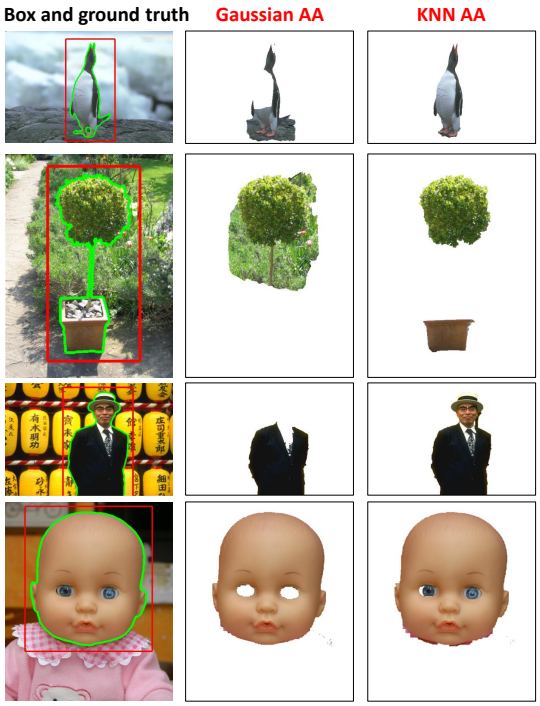


Fig. 8: Representative interactive segmentation results. Regularized average association (AA) with fixed bandwidth kernel (1) or adaptive KNN kernels (Example 3) is optimized as in [36]. Red boxes define initial clustering, green contours define ground-truth clustering. Table 1 provides the error statistics. Breiman's bias manifests itself by isolating the most frequent color from the rest.

regularization (boundary smoothness)	**average error, %** Gaussian AA	Gaussian NC	*KNN* AA	*KNN* NC
none†	20.4	17.6	**12.2**	12.4
Euclidean length*	15.1	16.0	**10.2**	11.0
contrast-sensitive*	9.7	13.8	**7.1**	7.8

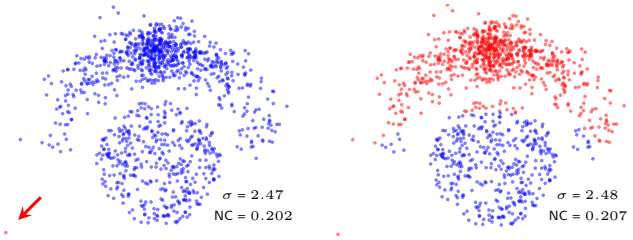
TABLE 1: Interactive segmentation errors. AA stands for the average association, NC stands for the normalized cut. Errors are averaged over the GrabCut dataset[37], see samples in Figure 8. *We use [36], [38] for a combination of Kernel K-means objective (8) with *Markov Random Field* (MRF) regularization terms. The relative weight of the MRF terms is chosen to minimize the average error on the dataset. †Without the MRF term, [36] and [38] correspond to the standard kernel K-means [8], [10].

are used to motivate *normalized cut* (NC) criterion (10) aiming at balanced clustering without "clumping" or "splitting".

We do not obeserve any evidence of the *mode isolation bias* in NC. However, Section 5.1 demonstrates that NC still has a bias to isolating sparse subsets. Moreover, using the general density analysis approach introduced in Section 4.2 we also show in Section 5.2 that *normalization* implicitly corresponds to some density-inverting embedding of the data. Thus, *mode isolation* (Breiman's bias) in this implicit embedding corresponds to the *sparse subset bias* of NC in the original data.

5.1 Sparse subset bias in Normalized Cut

The normalization in NC does not fully remove the bias to small isolated subsets and it is easy to find examples of "splitting" for weakly connected nodes, see Figure 9(a). The motivation



(a) NC for smaller bandwidth (bias to "sparsest" subsets) (b) NC for larger bandwidth (loss of non-linear separation)

Fig. 9: Normalized Cut with kernel (1) on the same data as in Figure 1(c,d). For small bandwidths NC shows bias to small isolated subsets (a). As bandwidth increases, the first non-trivial solution overcoming this bias (b) requires bandwidth large enough so that problems with non-linear separation become visible. Indeed, for larger bandwidths the node degrees become more uniform $d_p \approx const$ reducing NC to average association, which is known to degenerate into basic K-means (see Section 4.1). Thus, any further increase of σ leads to solutions even worse than (b). In this simple example no fixed σ leads NC to a good solution as in Figure 1(d). That good solution uses adaptive kernel from Section 4.3 making specific clustering criterion (AA, NC, or AC) irrelevant, see (68).

argument for the NC objective below Fig.1 in [1] implicitly assumes similarity matrices with zero diagonal, which excludes many common similarities like Gaussian kernel (1). Moreover, their argument is built specifically for an example with a single isolated point, while an isolated pair of points will have a near-zero NC cost even for zero diagonal similarities.

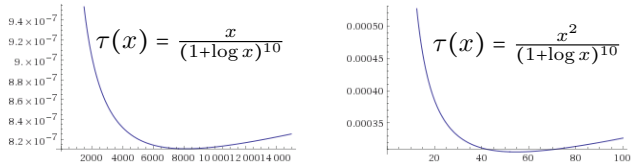
Intuitively, this NC issue can be interpreted as a bias to the "sparsest" subset (Figure 9a), the opposite of AA's bias to the "densest" subset, *i.e.* Breiman's bias (Figure 1c). The next subsection discusses the relation between these opposite biases in detail. In any case, both of these density inhomogeneity problems in NC and AA are directly addressed by our *density equalization* principle embodied in adaptive weights $w_p \propto 1/\rho_p$ in Section 3 or in the locally adaptive kernels derived in Section 4.3. Indeed, the result in Figure 1(d) can be replicated with NC using such adaptive kernel. Interestingly, [12] observed another data non-homogeneity problem in NC different from the sparse subset bias in Figure 9(a), but suggested a similar adaptive kernel as a heuristic solving it.

5.2 Normalization as density inversion

The bias to sparse clusters in NC with small bandwidths (Figure 9a) seems the opposite of mode isolation in AA (Figure 1c). Here we show that this observation is not a coincidence since NC can be reduced to AA after some density-inverting data transformation. While it is known [17], [8] that NC is equivalent to *weighted* kernel K-means (*i.e. weighted* AA) with some modified affinity, this section relates such kernel modification to an implicit density-inverting embedding where *mode isolation* (Breiman's bias) corresponds to *sparse clusters* in the original data.

First, consider standard weighted AA objective for any given affinity/kernel matrix $\hat{A}_{pq} = k(f_p, f_q)$ as in (42)

$$-\sum_k \frac{\sum_{pq\in S_k} w_p w_q \hat{A}_{pq}}{\sum_{p\in S_k} w_p}.$$



(a) density transform (65) (kernel normalization only) (b) density transform (66) (with additional point weighting)

Fig. 10: "Density inversion" in sparse regions. Using node degree approximation $d_p \propto \rho_p$ (67) we show representative density transformation plots (a) $\bar{\rho}_p = \tau(\rho_p)$ and (b) $\rho'_p = \tau(\rho_p)$ corresponding to AA with kernel modification $\hat{A}_{pq} = \frac{A_{pq}}{d_p d_q}$ (65) and additional point weighting $w_p = d_p$ (66) exactly corresponding to NC. This additional weighting weakens the density inversion in (b) compared to (a), see the x-axis scale difference. However, it is easy to check that the minima in (65) and (66) are achieved at some x^* exponentially growing with $\bar{N}$. This makes the density inversion significant for NC since $\bar{N}$ may equal the data size.

Clearly, weights based on node degrees $w = d$ and "normalized" affinities $\hat{A}_{pq} = \frac{A_{pq}}{d_p d_q}$ turn this into NC objective (10). Thus, average association (9) becomes NC (10) after two modifications:

- replacing A_{pq} by normalized affinities $\hat{A}_{pq} = \frac{A_{pq}}{d_p d_q}$ and
- introducing point weights $w_p = d_p$.

Both of these modifications of AA can be presented as implicit data transformations modifying denisty. In particular, we show that the first one "inverses" density turning sparser regions into denser ones, see Figure 10(a). The second data modification is generally discussed as a density transform in (43). We show that node degree weights $w_p = d_p$ do not remove the "density inversion".

For simplicity, assume standard Gaussian kernel (1) based on Euclidean distances $d_{pq} = \|f_p - f_q\|$ in $\mathcal{R}^N$

$$A_{pq} = \exp \frac{-d_{pq}^2}{2\sigma^2}.$$

To convert AA into NC we first need an affinity "normalization"

$$\hat{A}_{pq} = \frac{A_{pq}}{d_p d_q} = \exp \frac{-d_{pq}^2 - 2\sigma^2 \log(d_p d_q)}{2\sigma^2} = \exp \frac{-\hat{d}_{pq}^2}{2\sigma^2} \qquad (62)$$

equivalently formulated as a modification of distances

$$\hat{d}_{pq}^2 \;:=\; d_{pq}^2 + 2\sigma^2 \log(d_p d_q). \qquad (63)$$

Using a general approach in the proof of Theorem 3, there exists some Euclidean embedding $\bar{f}_p \in \mathcal{R}^{\bar{N}}$ and constant $h \geq 0$ such that

$$\bar{d}_{pq}^2 \;:=\; \|\bar{f}_p - \bar{f}_q\|^2 \;=\; \hat{d}_{pq}^2 + h^2[p \neq q]. \qquad (64)$$

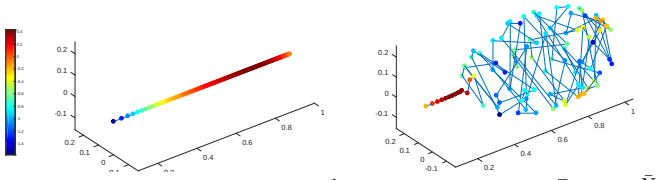
Thus, modified affinities $\hat{A}_{pq}$ in (62) correspond to the Gaussian kernel for the new embedding $\{\bar{f}_p\}$ in $\mathcal{R}^{\bar{N}}$

$$\hat{A}_{pq} \;\propto\; \exp \frac{-\bar{d}_{pq}^2}{2\sigma^2} \;\equiv\; \exp \frac{-\|\bar{f}_p - \bar{f}_q\|^2}{2\sigma^2}.$$

Assuming $d_q \approx d_p$ for features f_q near f_p, equations (63) and (64) imply the following relation for such neighbors of f_p

$$\bar{d}_{pq}^2 \;\approx\; d_{pq}^2 + h^2 + 4\sigma^2 \log(d_p).$$

Then, similarly to the arguments in (56), a small ball of radius ε centered at f_p in $\mathcal{R}^N$ and a ball of radius $\sqrt{\varepsilon^2 + h^2 + 4\sigma^2 \log(d_p)}$



(a) original data $\{f_p\} \subset \mathcal{R}^1$ (b) embedding $\{\bar{f}_p\} \subset \mathcal{R}^{\bar{N}}$

Fig. 11: Illustration of "density inversion" for 1D data. The original data points (a) are getting progressively denser along the line. The points are color-coded according to the log of their density. Plot (b) shows 3D approximation $\{y_p\} \subset \mathcal{R}^3$ of high-dimensional Euclidean embedding $\{\bar{f}_p\} \subset \mathcal{R}^{\bar{N}}$ minimizing metric errors $\sum_{pq}(\hat{d}^2_{pq} - \|y_p - y_q\|^2)^2$ where $\hat{d}_{pq}$ are distances (63).

at $\bar{f}_p$ in $\mathcal{R}^{\bar{N}}$ contain the same number of points. Thus, similarly to (57) we get a relation between densities at points f_p and $\bar{f}_p$

$$\bar{\rho}_p \;\approx\; \frac{\rho_p\, \varepsilon^N}{(\varepsilon^2 + h^2 + 4\sigma^2 \log(d_p))^{\bar{N}/2}}. \tag{65}$$

This implicit density transformation is shown in Figure 10(a). Sub-linearity in dense regions addresses mode isolation (Breiman's bias). However, sparser regions become relatively dense and kernel-modified AA may split them. Indeed, the result in Figure 9(a) can be obtained by AA with normalized affinity $\frac{A_{pq}}{d_p d_q}$.

The second required modification of AA introduces point weights $w_p = d_p$. It has an obvious equivalent formulation via data points replication discussed in Section 3, see Figure 4(a). Following (43), we obtain its implicit density modification effect $\rho'_p = d_p \bar{\rho}_p$. Combining this with density transformation (65) implied by affinity normalization $\frac{A_{pq}}{d_p d_q}$, we obtain the following density transformation effect corresponding to NC, see Figure 10(b),

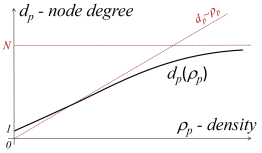
$$\rho'_p \;\approx\; \frac{d_p\, \rho_p\, \varepsilon^N}{(\varepsilon^2 + h^2 + 4\sigma^2 \log(d_p))^{\bar{N}/2}}. \tag{66}$$

The density inversion in sparse regions relates NC's result in Figure 9(a) to Breiman's bias for embedding $\{\bar{f}_p\}$ in $\mathcal{R}^{\bar{N}}$.

Figure 10 shows representative plots for density transformations (65), (66) using the following node degree approximation based on Parzen approach (11) for Gaussian affinity (kernel) A

$$d_p = \sum_q A_{pq} \propto \rho_p. \tag{67}$$

Empirical relation between d_p and ρ_p is illustrated below: some overestimation occurs for sparcer regions and underestimation happens for denser regions. The node degree for Gaussian kernels has to be at least 1 (for an isolated node) and at most N (for a dense graph).



6 Discussion (kernel clustering equivalence)

Density equalization with adaptive weights in Section 3 or adaptive kernels in Section 4 are useful for either AA or NC due to their density biases (mode isolation or sparse subset). Interestingly, kernel clustering criteria discussed in [1] such as normalized cut (NC), *average cut* (AC), average association (AA) or kernel K-means are practically equivalent for such adaptive methods. This can be seen both empirically (Table 1) and conceptually. Note, weights $w_p \propto 1/\rho_p$ in Section 3 produce modified data with near constant node degrees $d'_p \propto \rho'_p \propto 1$, see (67) and (43). Alternatively, KNN kernel (Example 3) with density equalizing bandwidth (61) also produce nearly constant node degrees $d_p \approx K$ where K is the neighborhood size. Therefore, both cases give

$$-\frac{\sum_{pq\in S^k} A_{pq}}{\sum_{p\in S^k} d_p} \;\propto\; -\frac{\sum_{pq\in S^k} A_{pq}}{K\,|S^k|} \;\overset{c}{\approx}\; \frac{\sum_{p\in S^k, q\in \bar{S}^k} A_{pq}}{K\,|S^k|}, \tag{68}$$

which correspond to NC (10), AA (9), and AC criteria. As discussed in [1], the last objective also has very close relations with standard partitioning concepts in spectral graph theory: *isoperimetric* or *Cheeger number*, *Cheeger set*, *ratio cut*.

This equivalence argument applies to the corresponding clustering objectives and is independent of specific optimization algorithms developed for them. Interestingly, the relation between (9) and basic K-means objective (3) suggests that standard Lloyd's algorithm can be used as a basic iterative approach for approximate optimization of all clustering criteria in (68). In practice, however, kernel K-means algorithm corresponding to the exact high-dimensional embedding $\{\phi_p\}$ in (3) is more sensitive to local minima compared to iterative K-means over approximate lower-dimensional embeddings based on PCA [14, Section 3.1][6].

Acknowledgements

The authors would like to thank Professor Kaleem Siddiqi (McGill University) for suggesting a potential link between Breiman's bias and the *dominant sets*. This work was generously supported by the Discovery and RTI programs of the National Science and Engineering Research Council of Canada (NSERC).

Appendix
Empirical density transformation plots

For adaptive bandwidths σ_p in (59) corresponding to any given $\tau(\rho)$, implicit quasi-isometric embedding in (50) allows to obtain an empirical scatter plot $\tau(\rho)_\Omega := \{(\rho'(f'_p), \rho(f_p)) \,|\, p \in \Omega\}$, *e.g.* shown in Figure 7(e), which could validate the "theoretical" plot $\rho' = \tau(\rho)$. Indeed, Parzen density (11) with kernel (1) for the new embedding $\{f'_p \,|\, p \in \Omega\} \subset \mathcal{R}^{N'}$ gives

$$\rho'(f'_p) \;\propto\; \sum_q e^{\frac{-\|f'_p - f'_q\|^2}{2}} \;\overset{c}{\propto}\; \sum_q e^{\frac{-\|f_p - f_q\|^2}{2\sigma_q^2}} \tag{69}$$

where $\overset{c}{\propto}$ is an equality up to some additive and multiplicative constants that follows directly from equation (50). Similar Parzen density estimate for the original embedding gives

$$\rho(f_p) \;\propto\; \sum_q e^{\frac{-\|f_p - f_q\|^2}{2}} \tag{70}$$

different from (69) only by the lack of σ_p in the exponent.

Note that adaptive bandwidths can also be used for density estimation often giving better accuracy. Interestingly, using adaptive σ_p for estimating density $\rho(f_p)$ leads to a variant of (70)

$$\rho(f_p) \;\propto\; \sum_q \frac{1}{\sigma_q^N} e^{\frac{-\|f_p - f_q\|^2}{2\sigma_q^2}} \tag{71}$$

6. K-means is also commonly used as a discretization heuristic for *spectral relaxation* [1] where a similar eigen analysis is motivated by spectral graph theory [40], [41], [42] defferently from PCA dimensionalty reduction in [14].

also different from (69). Now the difference is in the kernel *normalization*, which is required for density estimation in (71).

Note that *normalized* adaptive kernels is not a good idea for clustering. They turn $\mathcal{P}_\Sigma$ in probabilistic formulation (15) into a good approximation of the true data density. Thus, kernel K-means reduces to the continuous Gini criteria with Breiman's bias (Theorem 2). In contrast, *non-normalized* geodesic kernels (46) invalidate probabilistic formulation (15). Instead, they equalize the data density reducing Breiman's bias (see Sections 4.2 and 4.3).

REFERENCES

[1] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 2000. 1, 2, 3, 4, 6, 8, 9, 10

[2] L. Breiman, "Technical note: Some properties of splitting criteria," *Machine Learning*, vol. 24, no. 1, pp. 41–47, 1996. 1, 3, 4

[3] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural computation*, vol. 10, no. 5, pp. 1299–1319, 1998. 1

[4] V. Vapnik, *Statistical Learning Theory*. Wiley, 1998. 1, 2

[5] K. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Networks*, vol. 12, no. 2, pp. 181–201, 2001. 1

[6] R. Zhang and A. Rudnicky, "A large scale clustering scheme for kernel k-means," in *Pattern Recognition, 2002.*, vol. 4, 2002, pp. 289–292. 1

[7] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Trans. Neural Networks*, vol. 13, no. 3, pp. 780–784, 2002. 1, 3

[8] I. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means, spectral clustering and normalized cuts," in *KDD*, 2004. 1, 2, 8, 9

[9] M. Pavan and M. Pelillo, "Dominant sets and pairwise clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 167–172, 2007. 1, 5

[10] R. Chitta, R. Jin, T. Havens, and A. Jain, "Scalable kernel clustering: Approximate kernel k-means," in *KDD*, 2011, pp. 895–903. 1, 2, 8

[11] S. Jayasumana, R. Hartley, M. Salzmann, H. Li, and M. Harandi, "Kernel methods on Riemannian manifolds with Gaussian RBF kernels," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 37, no. 12, pp. 2464–2477, 2015. 1, 6

[12] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Advances in NIPS*, 2004, pp. 1601–1608. 1, 6, 7, 8, 9

[13] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2002. 2

[14] M. Tang, D. Marin, I. B. Ayed, and Y. Boykov, "Kernel Cuts: MRF meets kernel and spectral clustering," in *arXiv:1506.07439*, September 2016 (also submitted to IJCV). 2, 10

[15] V. Roth, J. Laub, M. Kawanabe, and J. Buhmann, "Optimal cluster preserving embedding of nonmetric proximity data," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1540—1551, 2003. 2, 7

[16] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007. 2

[17] F. Bach and M. Jordan, "Learning spectral clustering," *Advances in Neural Information Processing Systems*, vol. 16, pp. 305–312, 2003. 2, 9

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, August 2006. 3, 8

[19] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 1992. 3

[20] B. W. Silverman, *Density estimation for statistics and data analysis*. CRC press, 1986, vol. 26. 3

[21] A. J. Izenman, "Review papers: Recent developments in nonparametric density estimation," *Journal of the American Statistical Association*, vol. 86, no. 413, pp. 205–224, 1991. 3

[22] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. Wiley, 1973. 3, 5

[23] M. Kearns, Y. Mansour, and A. Ng, "An Information-Theoretic Analysis of Hard and Soft Assignment Methods for Clustering," in *Conf. on Uncertainty in Artificial Intelligence (UAI)*, August 1997. 3, 4, 6

[24] C. Fraley and A. E. Raftery, "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002. 3

[25] A. Criminisi and J. Shotton, *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013. 3

[26] Y. Boykov, H. Isack, C. Olsson, and I. B. Ayed, "Volumetric Bias in Segmentation and Reconstruction: Secrets and Solutions," in *International Conference on Computer Vision (ICCV)*, December 2015. 3, 6

[27] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984. 3

[28] T. S. Motzkin and E. G. Straus, "Maxima for graphs and a new proof of a theorem of turán," *Canad. J. Math*, vol. 17, no. 4, pp. 533–540, 1965. 5

[29] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International journal of computer vision*, vol. 42, no. 3, pp. 145–175, 2001. 5

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105. 5

[31] T. Cox and M. Cox, *Multidimensional scaling*. CRC Press, 2000. 5, 8

[32] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2006. 7

[33] J. A. Sethian, *Level set methods and fast marching methods*. Cambridge university press, 1999, vol. 3. 7

[34] J. Lingoes, "Some boundary conditions for a monotone analysis of symmetric matrices," *Psychometrika*, 1971. 7

[35] J. C. Gower and P. Legendre, "Metric and euclidean properties of dissimilarity coefficients," *Journal of classification*, vol. 3, no. 1, pp. 5–48, 1986. 7

[36] M. Tang, I. B. Ayed, D. Marin, and Y. Boykov, "Secrets of grabcut and kernel k-means," in *International Conference on Computer Vision (ICCV)*, Santiago, Chile, December 2015. 8

[37] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut - interactive foreground extraction using iterated graph cuts," in *ACM trans. on Graphics (SIGGRAPH)*, 2004. 8

[38] M. Tang, D. Marin, I. B. Ayed, and Y. Boykov, "Normalized Cut meets MRF," in *European Conference on Computer Vision (ECCV)*, 2016. 8

[39] Z. Wu and R. Leahy, "An optimal graph theoretic approach to data clustering: theory and its application to image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1101–1113, Nov 1993. 8

[40] J. Cheeger, "A lower bound for the smallest eigenvalue of the laplacian," *Problems in Analysis, R.C. Gunning, ed.*, pp. 195–199, 1970. 10

[41] W. Donath and A. Hoffman, "Lower bounds for the partitioning of graphs," *IBM J. Research and Development*, pp. 420–425, 1973. 10

[42] M. Fiedler, "A property of eigenvectors of nonnegative symmetric matrices and its applications to graph theory," *Czech. Math. J.*, vol. 25, no. 100, pp. 619–633, 1975. 10