L^2 -norm transformation for improving k -means clustering

Finding a suitable model by range transformation for novel data analysis

Piyush Kumar Sharma¹ · Gary Holness²

Received: 1 December 2016 / Accepted: 28 March 2017 / Published online: 29 April 2017
© Springer International Publishing Switzerland 2017

Abstract In the age of increasingly pervasive sensing applications, measurement of unknown pattern phenomena resulting in novel data presents a challenge to selection of appropriate modeling tools. Because there is no rich history of domain knowledge, one can easily make early commitments to poor modeling choices. Data transformation, a solution in effort to modify the data's geometry, can make important regularities more clear. The wrong transformation can damage the very pattern information one seeks to identify. In contrast to data transformation, we contribute an alternative method, range transformation focusing on altering the measurement tool. As a function, a model maps data inputs to a range. Focusing on transformations of the model's range, we can find a generally applicable way to alter the model's properties to best suit the data. Every modification to a function class, something we call *editing the function*, results in a change to the original function's range. This work contributes a method for modifying a broad class of models to suit novel data through range transformation. We investigate range transformation for a class of information theoretic transformations and evaluate impact on classifica-

tion and clustering. We also develop an optimization-based framework employing range transformation based on desired geometric properties and use it to improve a widely used model, k -means clustering.

Keywords Chisini–Jensen–Shannon divergence · CJS kernel · k -Means clustering · L^2 -norm · Range transformation · LIBS

1 Introduction

A number of successful methods in machine learning, most notably kernel methods, modify the input representation by a transformation $\phi : X \rightarrow Z$ that maps the input representation X into a new higher dimensional feature space representation Z . In doing so, models are afforded the ability to define complex nonlinear surfaces without paying increased cost in terms of learning complexity. In transforming the input representation, a singular transformation is applied across all regions of input space (Fig. 1).

In this work, we study a method that strives to transform the model's *measurement instrument* or distance function as an alternative to transformation of the input representation. The impact of this approach is a transformation of the range associated with the learned hypothesis. This results in an effective repositioning, thus changing their geometry, of data points in the input representation when learning a model. In transforming the range, more fine-grained control can be achieved over regions of the input representation beyond a singular transformation en masse of the input space itself. Consider a data set where points near the boundary between clusters are tightly packed. A solution for improved clustering would amplify distances corresponding to regions of input space near cluster boundaries and attenuate distances

This paper is an extension version of the DSAA-2016 paper: *Dilation of Chisini–Jensen–Shannon divergence* [39].

This work was supported by National Science Foundation Award# HRD-1242067 and Award# CNS-1205426.

✉ Piyush Kumar Sharma
pksharma12@students.desu.edu
Gary Holness
gholness@desu.edu

¹ Department of Mathematical Sciences, Delaware State University, Dover, DE 19904, USA

² Department of Computer and Information Sciences, Delaware State University, Dover, DE 19904, USA

Fig. 1 Schematic illustrates a model's range transformation

Fig. 2 Schematic illustrates our strategy of finding a suitable model by range transformation in both ways

corresponding to regions of input space further away from cluster boundaries. The effect of such a range transformation would make clusters more compact and well separated thus resulting in better clustering performance.

A range transformation is achieved two ways; the first changes or *edits* the model resulting in a transformation or geometric change to its range (what we call *the forward direction*), and the second begins with a desired geometric change to the range and rearranges or edits the model to achieve it (what we call *the reverse direction*) as shown in Fig. 2. An approach for range transformation in *the reverse direction* is the primary contribution of our work. We explore range transformation as a concept using two different models. For the forward direction, we explore a recently introduced information theoretic measure, CJSD [40]. For the reverse direction, we explore the very popular *k*-means clustering model.

The literature is full of examples from many disciplines where a model's performance is improved through its direct modification [14,16,33,50]. Recently published results in information theoretic models [41] demonstrated that Jensen–Shannon divergence (JSD) [10,21,24–26,38] does not provide adequate separation for drawing distinctions between subtly different distributions. This work reformulated (or edited) JSD with alternate operators in Fig. 3, namely, Chisini means [8]. The resulting Chisini–Jensen–Shannon divergences (CJSDs) rescale JSD's range. While the range remains uncountably infinite, it is transformed from the interval $[0, 1]$ to a different scale on the real number line to $[0, \infty]$ (Fig. 4).

Because of the broad applicability of information theoretic measures in describing regularities in data, for range transformation in the *forward direction* we focus on the development of range transformations arising from using Chisini means for function edits of JSDs (i.e., CJSDs). We derive the governing parameter, namely *dilation parameter*,

Fig. 3 Schematic presents model shaping by function edit for JSD

Fig. 4 CJSDs map input distributions onto nonnegative real number line

describing the relationship between ranges associated with different CJSDs (JSD edits). Closely related is the Bregman divergence framework which implements a distortion function (rearrangement of points under the function) [30]. Bregman divergence is limited to a specific set of functions; namely, those with a convex set domain and function f is strictly convex, i.e., its derivative is monotonically increasing [1]. Transformation of a function's range extends beyond the more limited set of functions required by Bregman divergence.

Extending beyond the recently introduced CJSDs, our work contributes an exploration of the mechanism responsible for CJSDs' improvement. Our larger contribution to machine learning is the development of range transformation as an approach to selection of suitable models. This is particularly important when analyzing novel data for which there is little knowledge about pattern phenomena (e.g., quantum effects due to photon energy absorption) or the resolution needed between model outputs in order to discriminate between classes. A method generally focused on resolution needed for discriminability can be valuable, particularly when there are subtle differences between class boundaries or members of different clusters. Using range transformation, one begins with the requisite resolution in the distance measure needed for discriminability, and then selects a transformation that achieves it. This approach opens a more broad class of functions available for modeling.

Open questions that we address concern: How does a CJSD transform a JSD's range? What are the properties these transformations? For what types of data are these reconstructions useful? How are distances rescaled for different CJSDs and what is the magnitude of this rescaling? What is the relationship between different CJSDs?

The range of CJSDs falls on the nonnegative real number line (Fig. 4). We describe *dilation mapping*, the rescaling of CJSDs in Fig. 9. Apart from rescaling, the dilation parameter also plays an important role in another set of parameters which we call *dilation ratios*, the relative dilation between two CJSDs. The intuition behind dilation ratios concerns the relative difference in scaling. A dilation ratio describes the rescaling magnitude. Our concept of CJSD dilation ratios can be considered similar to the ratio of magnification in Geometry [10]. Dilation ratios have interesting properties, that is they can be put in rank order. With this one can begin to surmise the improvement in discrimination between input distributions afforded by the family of CJSDs.

Motivated by applications in laser spectroscopy, the original CJSD contribution [41] applied CJSD in a new kernel for SVM classification. Our experiments carry forward investigation of CJSDs' utility beyond laser spectroscopy with validation on a synthetic data set where subtlety between the classes is controlled. This validation tests the applicability of CJSDs on arbitrary data characterized by the general properties of stochasticity and subtlety. Most importantly, our experiments empirically confirm our theories concerning dilation parameter.

We also study range transformation in the *backwards direction* in Fig. 2. There are many examples in the literature for improving the performance of a model by transformation on data. For example, k -means assumes clusters are spherical in shape. This technique works well for a certain types of data sets; however, it performs poorly for others. This stems from trying to construct clusters comprised of equal numbers of data points. Examples in the literature demonstrate that this problem can be solved by data transformation (such as polar coordinate transformation). With range transformation, there is no early commitment about the data's geometry and because data are not projected into a higher dimensional representation, we do not run against the curse of dimensionality. Statistical learning theory tells us that one can only increase dimensionality to the point where the sample size remains large enough to pay for the increased complexity [49].

We implement range transformation in the backward direction using an *optimization framework* to modify the model. The contributed implementation of range transformation automates alteration of the measurement tool, namely the distance measure. We chose L^2 -norm because of its ubiquity. Moreover, the L^2 -norm appears in many machine learning algorithms—several SVM kernels (RBF, CJSDs, Gaussian,

String, etc.), nonlinear manifold embeddings (ISOMAP, LLE, Laplacian eigenmaps, etc.), and clustering (k -means, hierarchical, mean shift, etc.). Consider k -means clustering. In Sect. 7.2, we demonstrate that range transformation of L^2 -norm gives improved performance on novel spectroscopy data.

2 Chisini–Jensen–Shannon divergences

Jensen–Shannon divergence (JSD) is a symmetric and smoothed version of Kullback–Leibler divergence (KLD) [25, 26]. Let, $P=\{p_i\}_{i=1}^N$, and $Q=\{q_i\}_{i=1}^N$ be two probability distributions, where p_i and q_i are the respective probabilities associated with the i th state (possible values). Jensen–Shannon divergence (JSD) of Q from P is

$$\text{JSD}(P||Q) = \frac{1}{2} \left[\sum_{i=1}^N p_i \log \frac{p_i}{M_i} + \sum_{i=1}^N q_i \log \frac{q_i}{M_i} \right] \quad (1)$$

$$\text{where } M \text{ is Arithmetic Mean, } \text{A.M.} = \left\{ \frac{p_i + q_i}{2} \right\}_{i=1}^N \quad (2)$$

Average value of M in (2) computes JSD precisely at the midpoint of P and Q ; however, when distributions have subtle differences, it is hard to distinguish them. Therefore, reformulation of JSD for better discrimination with alternate operators was proposed [41]. Though authors demonstrated the utility of this reformulation by replacing (2) by (3) or (4), given as follows:

$$\text{Geometric Mean, G.M.} = \left\{ \sqrt{p_i q_i} \right\}_{i=1}^N \quad (3)$$

$$\text{Harmonic Mean, H.M.} = \left\{ \frac{2 p_i q_i}{p_i + q_i} \right\}_{i=1}^N \quad (4)$$

this reformulation is *not limited* to Chisini mean.

3 Transformations

An important question concerns the type of transformations one might perform on a function's range. We investigate different types of transformations which are useful in disciplines, such as, mechanics, engineering, geometry, and computer graphics. These transformations are well known for their applicability in the rigid and nonrigid body deformation and distortion. In computer graphics, their applications are also well studied in modeling the shape of an object. We consider the effect of different transformations to understand how a function's range can be modified. This helps provide an answer to the questions *How does a CJSD transform a JSD's range?* and *What are the properties of these transformations?*

Fig. 5 Some affine transformations

3.1 Affine transformation

Affine transformation is a linear mapping that preserves collinearity. After transformation, all points lying on a line initially still lie on a line. An affine transformation is not an angle or length preserving transformation. However, it does preserve proportions on lines. Moreover, it preserves ratios of distances so that sets of parallel lines remain parallel after an affine transformation. Therefore, a midpoint on a straight line will remain a midpoint after transformation. Affine transformations do not move any object from the affine space R^3 to the plane at infinity or conversely. They are considered a type of projective transformations and also known as affinity [42].

Affine transformations are widely used in geometry to correct deformations or distortions caused by camera angles. In image registration satellite camera cause wide angel lens distortion. Affine transformations are used to correct such distortions. This is sometimes achieved by transforming and fusing the images to a large, flat coordinate system. Some affine transformations are *translation*, *scaling*, *rotation*, *shear*, *reflection*, *expansion*, *contraction*, *composition*, *homothety*, *dilation*, and *spiral similarity* shown in Fig. 5.

3.2 Geometrical deformation

Deformation is a function from $R^3 \rightarrow R^3$ which maps each point $P(x, y, z)$ in Euclidean space to a point $P(x', y', z')$. As demonstrated for affine transformation in previous subsection, a desired geometric deformation can be obtained by function compositions. Therefore, it is a transformations from some initial to some final geometry.

Deformation can be defined as a change in shape, volume, position or rotation due to an applied stress. In case of rigid body, such changes can be obtained by affine transformation we discussed in previous subsection; however, in case

Fig. 6 Homogeneous and heterogeneous deformation

Fig. 7 A change occurs in the point cloud position after deformation. An injective mapping exists from initial to final position

of nonrigid body it is called *distortion* and can be obtained by *dilation*. Under such transformations, we can obtain *homogeneous* and *heterogeneous* deformations (Fig. 6).

Homogeneous deformation is obtained by uniformly applying a transformation on a rigid body. It preserves parallelism of lines, however, each geometrical object is distorted into a different shape, a cube into a prism, a square into a parallelograms, a circle into an ellipse, and a sphere into an ellipsoid.

Heterogeneous deformation is obtained by irregularly or nonuniformly applying a transformation on a rigid body. It does not preserve parallelism of lines, and each object, such as, square, cube, and circle is distorted into an irregular complex shapes.

Each particle on a geometrical object has its unique location in space. As these locations change under deformation, we may assert that there exists a mapping between initial and final positions as shown in Fig. 7. Let, vectors $\mathbf{x}$ and $\mathbf{y}$ represent the initial and the final position, respectively, of a particular particle on interest on any geometrical shape. Because no two distinct particles can be deformed into the same location, this mapping is an injective mapping defined by $\mathbf{y} = f(\mathbf{x})$, where f is the mapping function [4].

In our results we prove that there exist a *dilation mapping* between Chisini–Jensen–Shannon divergences (CJSDs). These CJSDs are resulted by the transformation of JSD using dilation. Our concept of range transformation of CJSDs is similar to the above concept of mapping for geometric deformations. Such mapping uniquely defines the positions of particles (or inputs to CJSDs) before and after deformation (range transformation in CJSDs). Controlling the governing parameter(s) responsible for deformation (range transforma-

tion for CJSDs), we can preset the final positions of particles. By this approach, we can implement the desired relative spacing between two CJSDs; thus, a desired magnification by scaling can be obtained. Our approach, JSD's range transformation, is particularly helpful with nuanced data sets where it is difficult to distinguish classes from each other. This modification gives rise to the formulation of CJSD-based kernels which improve generalization performance in SVM classification tasks over our JSD-based kernel (kernel without dilation) Sect. 6.2.

3.3 Modulation and perturbation transformation

Deformation techniques interactively model a rough surface. Because some objects have a geometric representation, these techniques study the point location of a deformed object which are useful in modeling [3, 7, 37]. In the area of computer graphics, two known deformation-based transformations techniques are

Modulation Transformation, T_M , for modulation M and amplitude vector (A_x, A_y, A_z) is defined as:

$$T_M(x, y, z) = (x + tA_x, y + tA_y, z + tA_z), \quad t = M(x, y, z)$$

Perturbation Transformation, T_P , for perturbation P and amplitude vector A is defined as:

$$T_P(x, y, z) = (x + uA, y + vA, z + wA), \\ (u, v, w) = P(x, y, z)$$

These transformations can introduce a desired irregularity or deformation to an object at the right position, orientation and size. Therefore, the proper tuning of the parameter values can provide a wide variety of visual effects. Moreover, these deformations can be reformulated by combining them with other affine transformation mentioned earlier.

3.4 Isotropic versus non-isotropic dilation

An *isotropic dilation* is a uniform contraction or expansion of the plane or space about some fixed point (center) as illustrated in Fig. 5. If this point is the origin, the transformation is homogeneous isotropic dilation. Even though it is an affine transformation, it is not an isometry. However, the product of such a dilation and isometry is a similarity, which is a one-to-one mapping. In similarity mapping each length is multiplied by a same number which is called *similarity ratio*, or *strain* in physics [28] (For CJSDs we called it *dilation ratios* Sect. 4.1).

An *anisotropic dilation* is a contraction or expansion of the plane or space whose ratio is dependent on orientation. Under this transformation, points move either far away from a center or closer to a center by unequal distances. There-

Fig. 8 (Left) points A and B in 2D plane under anisotropic dilation with scale factors k_1 and k_2 . (Right) obtained elliptical shape after anisotropic dilation

fore, an object is changed nonuniformly which may appear as a distortion or deformation in original object's shape. For example, an isosceles or an equilateral triangle remains no longer an isosceles or an equilateral triangle under this transformation.

Let, in 3D case, point (x, y, z) is transformed to (x', y', z') . Let, X be a vector in Cartesian coordinate system. In matrix form, the transformation matrix can be represented by $X^T = KX$, where K is the transformation matrix. The transformation matrices (*dilation matrices*) for isotropic and non-isotropic dilations are given as:

$$K_{\text{Isotropic}} = \begin{bmatrix} k & 0 & 0 & 0 \\ 0 & k & 0 & 0 \\ 0 & 0 & k & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$$K_{\text{Non-Isotropic}} = \begin{bmatrix} k_1 & 0 & 0 & 0 \\ 0 & k_2 & 0 & 0 \\ 0 & 0 & k_3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Here, for isotropic case, $x' = kx$, $y' = ky$, $z' = kz$, and for non-isotropic case, $x' = k_1x$, $y' = k_2y$, $z' = k_3z$.

It was shown that two points in a 2D plane produce an elliptical shape under non-isotropic dilation [28]. Let, $A \equiv (x_1, y_1)$ and $B \equiv (x_2, y_2)$ be the two points in 2D plane under non-isotropic dilation transformation by scale factors k_1 and k_2 toward X and Y axis, respectively, Fig. 8. Let, the line on which these two points lie, makes an angle ϕ from the X axis.

Distance between A and B : $d_1 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$

Distance after dilation: $d_2 = \sqrt{(x'_1 - y'_1)^2 + (x'_2 - y'_2)^2}$

But, $x' = k_1x$ and $y' = k_2y$ as defined above for anisotropic dilation. Therefore,

$$k_\phi = \frac{d_1}{d_2} = \frac{\sqrt{k_1^2(x_1 - y_1)^2 + k_2^2(x_2 - y_2)^2}}{\sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}}$$

Substituting $y = x \tan \phi$ in previous expression, we get,

$$k_\phi = \sqrt{k_1^2 \cos^2 \phi + k_2^2 \sin^2 \phi}$$

We can control the amount of dilation by varying k_1 and k_2 toward X and Y axis, respectively. This means that k_ϕ and ϕ represent an ellipse with k_1 and k_2 as their major or minor axes.

Diagonal elements of K are responsible for dilation. In such case, transformation is uniform as each point or distance is dilated by the same number k . This is very useful as we can transform things from microscale to macroscale. Its utility comes handy in application where micro-level things needs to be measured with exact precision, and measuring equipments are limited in their measuring capability. For example, bio-chemists depend heavily on high precision microscope and sensors for the measurement of distances between bio-macro molecules (protein, RNA, etc.). In such domain, distances are measured at nanometer scale level. Magnification of relative distributions of distance (e.g., some spectroscopic technique for measurement) by isotropic dilation can give a better resolution of chemical structure bio-macro molecules. Moreover, astronomical distances are extremely large, measured in light years (relative distances between heavenly bodies). Again, such measurements depend heavily on advanced equipments (satellite, telescope, etc.). Dilation can be used to scale down these measurements.

4 Dilation of CJSDs

Definition 1 *Dilation* of a function means either stretching away from an axis or compressing toward an axis.

Definition 2 There exists a bijection mapping from domain of input distributions to their respective ranges mapped by CJSDs. We call this function mapping a *dilation mapping* of CJSDs.

Consider S a set of pairs s_i of probability distributions $P = \{p_i\}$ and $Q = \{q_i\}$, where, $i = 1, \dots, n$, and $p_i, q_i \in [0, 1] \in \mathbb{R}$. So, there are n such pairs of probability distributions in set S , and for each pair we compute their JSDs.

Let, the functions JSD_{AM} , JSD_{GM} , and JSD_{HM} map each $s_i \in S$ to images $x_i \in X$, $y_i \in Y$, and $z_i \in Z$, respectively, as follows (Fig. 9):

$$\text{JSD}_{\text{AM}} : S \mapsto X$$

$$\text{JSD}_{\text{GM}} : S \mapsto Y$$

$$\text{JSD}_{\text{HM}} : S \mapsto Z$$

JSD with subscripts AM, GM, and HM refer to the respective divergences reformulated with Chisini mean (arithmetic, geometric, and harmonic) Fig. 17.

Fig. 9 Schematic of dilation mappings for CJSDs

4.1 Dilation ratios

The following inequality was proved in [41]:

$$\text{JSD}_{\text{AM}} < \text{JSD}_{\text{GM}} < \text{JSD}_{\text{HM}} \quad (5)$$

Consider the ratios:

$$\alpha = \frac{\text{JSD}_{\text{GM}}}{\text{JSD}_{\text{AM}}} > 1 \quad (6)$$

$$\beta = \frac{\text{JSD}_{\text{HM}}}{\text{JSD}_{\text{AM}}} > 1 \quad (7)$$

$$\gamma = \frac{\text{JSD}_{\text{HM}}}{\text{JSD}_{\text{GM}}} > 1 \quad (8)$$

(6), (7), and (8) are true by (5). We call α , β , γ the *dilation ratios* which give us the scale factor that tell us the amount of magnification in scaling. Dilation ratios show how CJSDs space apart a given distributions w.r.t. each other.

4.2 Relationship between α , β , γ

$$\beta - \alpha = \frac{\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}}{\text{JSD}_{\text{AM}}} > 0 \implies \boxed{\beta > \alpha} \quad (9)$$

$$\text{and, } \frac{1}{\gamma} - \frac{1}{\beta} = \frac{\text{JSD}_{\text{GM}} - \text{JSD}_{\text{AM}}}{\text{JSD}_{\text{HM}}} > 0 \implies \boxed{\beta > \gamma} \quad (10)$$

Alternately, we can prove (10) by substituting (6) and (7) in (8), i.e.,

$$\begin{aligned}\gamma &= \frac{\beta \cdot \text{JSD}_{\text{AM}}}{\alpha \cdot \text{JSD}_{\text{AM}}} \\ \Rightarrow \boxed{\beta = \gamma\alpha}\end{aligned}\quad (11)$$

As $\alpha, \gamma > 1$, we have $\beta > \gamma$

$$\begin{aligned}\text{Now, } \gamma - \alpha &= \frac{\text{JSD}_{\text{HM}}}{\text{JSD}_{\text{GM}}} - \frac{\text{JSD}_{\text{GM}}}{\text{JSD}_{\text{AM}}} \\ &= \frac{\text{JSD}_{\text{AM}}\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}^2}{\text{JSD}_{\text{AM}}\text{JSD}_{\text{GM}}}\end{aligned}$$

As the numerator is less than 0 by Proposition. 2,

$$\Rightarrow \boxed{\gamma < \alpha}$$

$$\text{Also, } \gamma - \frac{1}{\alpha} = \frac{\text{JSD}_{\text{HM}} - \text{JSD}_{\text{AM}}}{\text{JSD}_{\text{GM}}} > 0 \quad (12)$$

$$\Rightarrow \boxed{\gamma > \frac{1}{\alpha}} \quad (13)$$

As $\alpha > 1 \Rightarrow \frac{1}{\alpha} = \gamma_+ =$ a nonnegative real number smaller than 1.

From (9), (12), and (13);

$$\Rightarrow \boxed{\beta > \alpha > \gamma > \frac{1}{\alpha}} \quad (14)$$

Proposition 1 Prove that CJSDs have a dilation relationship. Given arbitrary input S , JSD_{GM} 's scaling is greater than JSD_{AM} , likewise JSD_{HM} 's scaling is greater than both JSD_{AM} and JSD_{GM} .

Proof Define $x_i \in [0, 1]$, $y_i \in [0, \infty)$, and $z_i \in [0, \infty]$ in co-domains X , Y , and Z , respectively, as mapped images of each tuple s_i in S Fig. 17. For JSD_{AM} , s_i is mapped to co-domain X described by the interval $[0, 1] \in \mathbb{R}$, while JSD_{GM} and JSD_{HM} map s_i to co-domain Y and Z , respectively, described by the interval $[0, \infty]$. This means that JSD_{GM} and JSD_{HM} provide greater scaling than JSD_{AM} . Moreover, from (14), JSD_{HM} provides greater scaling than both JSD_{AM} and JSD_{GM} (Fig. 17). $\square$

Remark 1 Proposition 1 shows that the range $[0, \infty]$ of JSD_{GM} and JSD_{HM} is scaled differently. We show that they have the same cardinality. For any interval on the real number line and the real number line itself have the same cardinality, namely uncountably infinite. This is demonstrated by proof that a bijection exists between intervals on $\mathbb{R}$ and $\mathbb{R}$. Consider $[0, 1]$ from the set of intervals on $\mathbb{R}$. Map $[0, 1]$ to $[0, \infty)$ by taking an element $a \in [0, 1]$ and mapping it through $f(x) = \frac{1}{x}$ to element $b \in [0, \infty]$. Algebraically the value of $f(a) = \frac{1}{a}$. Therefore, $\frac{1}{a} = b$. Through the inverse mapping, b 's pre-image is $f^{-1}(b) = \frac{1}{b} = a$. This also applies

to arbitrary intervals $[i, j] \in \mathbb{R}$. Because an onto mapping exists and a is the only value mapped onto b , the pair (a, b) is a unique mapping. Hence, $[0, 1]$ and $[0, \infty)$ have the same cardinality.

Proposition 2 $\text{JSD}_{\text{AM}}\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}^2 < 0$.

Proof As (5) holds, we have the following:

$$\text{JSD}_{\text{GM}} - \text{JSD}_{\text{AM}} > 0 \text{ \& \> } \text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}} > 0$$

$$\therefore (\text{JSD}_{\text{GM}} - \text{JSD}_{\text{AM}})(\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}) > 0$$

$$\begin{aligned}\text{or } \text{JSD}_{\text{GM}}\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}^2 - \text{JSD}_{\text{AM}}\text{JSD}_{\text{HM}} \\ + \text{JSD}_{\text{AM}}\text{JSD}_{\text{GM}} > 0\end{aligned}$$

$$\begin{aligned}\text{or } \text{JSD}_{\text{GM}}^2 + \text{JSD}_{\text{AM}}\text{JSD}_{\text{HM}} < \text{JSD}_{\text{AM}}\text{JSD}_{\text{GM}} \\ + \text{JSD}_{\text{GM}}\text{JSD}_{\text{HM}}\end{aligned}$$

Adding $-2\text{JSD}_{\text{GM}}^2$ on both sides:

$$\text{JSD}_{\text{AM}}\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}^2 < \text{JSD}_{\text{AM}}\text{JSD}_{\text{GM}}$$

$$+ \text{JSD}_{\text{GM}}\text{JSD}_{\text{HM}} - 2\text{JSD}_{\text{GM}}^2$$

$$< \text{JSD}_{\text{GM}}(\text{JSD}_{\text{AM}} + \text{JSD}_{\text{HM}} - 2\text{JSD}_{\text{GM}})$$

$$\text{Now, } \text{JSD}_{\text{GM}} = \frac{1}{2}[\text{JSD}_{\text{AM}} + \text{JSD}_{\text{HM}}]$$

[Proved in [41]]

$$\Rightarrow \boxed{\text{JSD}_{\text{AM}}\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}^2 < 0}$$

$\square$

Proposition 3 Prove the following equality,

$$\begin{aligned}\text{JSD}_{\text{HM}} - \text{JSD}_{\text{AM}} &= 2(\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}) \\ &= 2(\text{JSD}_{\text{GM}} - \text{JSD}_{\text{AM}})\end{aligned}$$

Proof $\text{JSD}_{\text{HM}} - \text{JSD}_{\text{AM}} =$

$$\begin{aligned}&\frac{1}{2} \left[\left\{ \sum_{i=1}^n p_i \log \left(\frac{p_i}{\text{HM}} \right) + \sum_{i=1}^n q_i \log \left(\frac{q_i}{\text{HM}} \right) \right\} \right. \\ &\quad \left. - \left\{ \sum_{i=1}^n p_i \log \left(\frac{p_i}{\text{AM}} \right) + \sum_{i=1}^n q_i \log \left(\frac{q_i}{\text{AM}} \right) \right\} \right] \\ &= \frac{1}{2} \left[\sum_{i=1}^n p_i \log \left(\frac{\text{AM}}{\text{HM}} \right) + \sum_{i=1}^n q_i \log \left(\frac{\text{AM}}{\text{HM}} \right) \right] \\ &= \frac{1}{2} \sum_{i=1}^n \left[p_i \log \left(\frac{\text{AM}}{\frac{\text{GM}^2}{\text{AM}}} \right) + q_i \log \left(\frac{\text{AM}}{\frac{\text{GM}^2}{\text{AM}}} \right) \right] \\ &\quad [\because \text{G.M.}^2 = \text{A.M.} \times \text{H.M.}] \\ &= \frac{1}{2} \left[\sum_{i=1}^n p_i \log \left(\frac{\text{AM}}{\text{GM}} \right)^2 + \sum_{i=1}^n q_i \log \left(\frac{\text{AM}}{\text{GM}} \right)^2 \right]\end{aligned}$$

Therefore, we get

$$\text{JSD}_{\text{HM}} - \text{JSD}_{\text{AM}} = \sum_{i=1}^n \left[(p_i + q_i) \log \left(\frac{\text{AM}}{\text{GM}} \right) \right] \quad (15)$$

Similarly, we get the following two equations:

$$\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}} = \frac{1}{2} \sum_{i=1}^n \left[(p_i + q_i) \log \left(\frac{\text{AM}}{\text{GM}} \right) \right] \quad (16)$$

$$\text{JSD}_{\text{GM}} - \text{JSD}_{\text{AM}} = \frac{1}{2} \sum_{i=1}^n \left[(p_i + q_i) \log \left(\frac{\text{AM}}{\text{GM}} \right) \right] \quad (17)$$

The right-hand side of (16) and (17) is also known as *arithmetic–geometric mean divergence* [20, 46]. Above three steps give the desired proof. $\square$

4.3 Order of dilation in CJSDs

The question that remains up to what order JSD_{GM} w.r.t. JSD_{AM} and JSD_{HM} w.r.t. JSD_{AM} & JSD_{GM} space apart distributions on the real number line? Representing the right-hand sides of (15) as κ . Equations (15), (16), and (17) can be written as:

$$\text{JSD}_{\text{HM}} = \text{JSD}_{\text{AM}} + \kappa \quad (18)$$

$$\text{JSD}_{\text{HM}} = \text{JSD}_{\text{GM}} + \frac{1}{2}\kappa \quad (19)$$

$$\text{JSD}_{\text{GM}} = \text{JSD}_{\text{AM}} + \frac{1}{2}\kappa \quad (20)$$

Remark 2 Since (18), (19), and (20) represent a consistent and dependent system of linear equations, it is not possible to solve it algebraically.

Substituting (20) in (6):

$$\alpha = 1 + \frac{1}{2} \left[\frac{\kappa}{\text{JSD}_{\text{AM}}} \right] \quad (21)$$

Substituting (18) in (7):

$$\beta = 1 + \left[\frac{\kappa}{\text{JSD}_{\text{AM}}} \right] \quad (22)$$

Substituting (19) in (8):

$$\gamma = 1 + \frac{1}{2} \left[\frac{\kappa}{\text{JSD}_{\text{GM}}} \right] \quad (23)$$

Substituting (20) in (23):

$$\therefore \gamma = 1 + \left[\frac{\kappa}{2\text{JSD}_{\text{AM}} + \kappa} \right] \quad (24)$$

Relative improvement (R.I.) in scaling is computed as:

$$\begin{aligned} & \frac{\text{JSD}_{\text{GM}} - \text{JSD}_{\text{AM}}}{\text{JSD}_{\text{AM}}} \times 100\% \\ & \frac{\text{JSD}_{\text{HM}} - \text{JSD}_{\text{GM}}}{\text{JSD}_{\text{GM}}} \times 100\% \\ & \frac{\text{JSD}_{\text{HM}} - \text{JSD}_{\text{AM}}}{\text{JSD}_{\text{AM}}} \times 100\% \end{aligned}$$

$$(\alpha - 1) \times 100\%, \quad (\beta - 1) \times 100\%, \quad (\gamma - 1) \times 100\%$$

[By (6), (7), and (8)]

Remark 3 While Eqs. (18), (19), and (20) compute the differences (spacing) between CJSDs, and Eqs. (21), (22), and (24) compute dilation ratios which tell us precisely up to what order (scale factor) CJSDs will space apart the given distributions. Moreover, Eqs. (21), (22), and (24) also confirm the validity of (14). By substituting expression for dilation ratios (21), (22), and (24) into (6), (7), and (8), respectively, we get expressions ordering dilations, i.e., (18), (19), (20).

Equations (21), (22), and (24) are advantaged over (6), (7), and (8) because they depend only on JSD_{AM} and κ . Thus, we can compute α , β , and γ without first computing JSD_{GM} and JSD_{HM} . Also, from Eqs. (18) and (20), we can compute JSD_{HM} and JSD_{GM} implicitly as they depend only on JSD_{AM} and κ . See the modified schematic of this approach in Fig. 10.

4.4 Alternate representation of CJSDs and κ

We call κ the **dilation parameter**. Elaborate κ in order to understand its principal constituents:

$$\begin{aligned} \kappa &= \sum_{i=1}^n \left[(p_i + q_i) \log \left(\frac{\text{AM}}{\text{GM}} \right) \right] \\ &= \sum_{i=1}^n \left[(p_i + q_i) \log \left(\frac{p_i + q_i}{2\sqrt{p_i q_i}} \right) \right] \\ &= \sum_{i=1}^n \left[(p_i + q_i) \log (p_i + q_i) - (p_i + q_i) \log 2 \right. \\ &\quad \left. - \frac{1}{2} p_i \log p_i - \frac{1}{2} p_i \log q_i - \frac{1}{2} q_i \log p_i - \frac{1}{2} q_i \log q_i \right] \end{aligned}$$

Fig. 10 Modified schematic of dilation mapping shown in Fig. 17. JSD_{HM} (set Z) can be computed implicitly by mappings ϕ_3, ϕ_2

Expanding the cross entropy terms:

$$\begin{aligned}
 &= \sum_{i=1}^n \left[(p_i + q_i) \log(p_i + q_i) - (p_i + q_i) \log 2 - \frac{1}{2} p_i \log p_i \right. \\
 &\quad \left. - \frac{1}{2} p_i \log p_i - \frac{1}{2} p_i \log \frac{q_i}{p_i} - \frac{1}{2} q_i \log q_i - \frac{1}{2} q_i \log \frac{p_i}{q_i} \right. \\
 &\quad \left. - \frac{1}{2} q_i \log q_i \right] \\
 &= \sum_{i=1}^n \left[(p_i + q_i) \log(p_i + q_i) - (p_i + q_i) \log 2 \right. \\
 &\quad \left. - p_i \log p_i + \left(\frac{1}{2} p_i \log \frac{p_i}{q_i} + \frac{1}{2} q_i \log \frac{q_i}{p_i} \right) - q_i \log q_i \right] \\
 &= \left[H_P + H_Q - H_{P+Q} - \sum_{i=1}^n (p_i + q_i) + \frac{1}{2} (SKLD) \right] \\
 &\quad \kappa = \left[I(P, Q) - \sum_{i=1}^n (p_i + q_i) + \frac{1}{2} (SKLD) \right] \quad (25)
 \end{aligned}$$

where H_P, H_Q , and H_{P+Q} are the Shannon entropies of the distributions P and Q , and their joint distribution $P + Q$, respectively. $I(P, Q)$ is the *mutual information* [9]. SKLD is the symmetric KL divergence. κ can be used for deriving alternate representation for JSD_{GM} and JSD_{HM} . First, expand JSD_{AM} :

$$\begin{aligned}
 JSD_{AM} &= \frac{1}{2} \sum_{i=1}^n \left[p_i \log \frac{p_i}{(p_i + q_i)/2} + q_i \log \frac{q_i}{(p_i + q_i)/2} \right] \\
 &= \frac{1}{2} \sum_{i=1}^n \left[p_i \log p_i - p_i \log \frac{(p_i + q_i)}{2} + q_i \log q_i \right. \\
 &\quad \left. - q_i \log \frac{(p_i + q_i)}{2} \right] \\
 &= \frac{1}{2} \sum_{i=1}^n \left[p_i \log p_i + q_i \log q_i - (p_i + q_i) \log(p_i + q_i) \right. \\
 &\quad \left. + (p_i + q_i) \log 2 \right]
 \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{2} \left[-H_P - H_Q + H_{P+Q} + \sum_{i=1}^n (p_i + q_i) \right] \\
 JSD_{AM} &= \frac{1}{2} \left[-I(P, Q) + \sum_{i=1}^n (p_i + q_i) \right] \quad (26)
 \end{aligned}$$

Substituting (25) and (26) in (20):

$$JSD_{GM} = \frac{1}{4} [SKLD] \quad (27)$$

Though relation (27) was established by [41], ours provides an alternate proof. Substituting (25) and (26) in (18):

$$\begin{aligned}
 JSD_{HM} &= \frac{1}{2} \left[H_P + H_Q - H_{P+Q} - \sum_{i=1}^n (p_i + q_i) + SKLD \right] \\
 JSD_{HM} &= \frac{1}{2} \left[I(P, Q) - \sum_{i=1}^n (p_i + q_i) + SKLD \right] \quad (28)
 \end{aligned}$$

Alternate representations in (25), (26), (27), and (28) are important in understanding the effect of κ 's spacing apart of JSD_{AM} outputs on the real number line. Section 4.2 discussed mutual differences between CJSDs. Parameter κ modifies JSD_{AM} 's values given by (18) and (20) by dilating them further apart on the real number line. Moreover, from (25) we can also tell precisely how much it will dilate. In order to get JSD_{GM} and JSD_{HM} , we dilate JSD_{AM} by a quantity $\kappa/2$, and κ , respectively.

4.5 Modified dilation mapping

We redefine *dilation mapping* defined earlier in Def. 2. Consider new dilation mapping functions ϕ_1, ϕ_2 , and ϕ_3 such that:

$$\phi_1 : X_{\kappa/2} \mapsto Y \quad \phi_2 : Y_{\kappa/2} \mapsto Z \quad \phi_3 : X_{\kappa} \mapsto Z$$

X_{κ} contains the corresponding elements of X and κ , $X_{\kappa/2}$ contains the corresponding elements of sets X and $\kappa/2$ and $Y_{\kappa/2}$ contains the corresponding elements of sets Y and $\kappa/2$ where, κ can be computed implicitly by (20), that is, $\kappa = 2(Y - X)$. In Fig. 10 mappings ϕ_1, ϕ_2 and ϕ_3 represent (20), (19), and (18), respectively.

5 L^2 -norm range transformation

So far we studied what happens to the range when model changes and proved it in the forward direction. Now, we will

study what happens to the model when range changes and prove it in the backward direction (Fig. 2). For experimental purpose, we choose L^2 -distance (Euclidean distance) which is commonly used in real-world applications, such as face recognition, financial prediction [17], and data mining [13].

L^2 -distance is widely used measure for computing distance between two points. Its range falls on the nonnegative real number line. Let, $X = \{x_1, x_2, \dots, x_n\}$ be a data set of length n . Then, for any two given data points $x_i, x_j \in X$, it is defined as:

$$L^2 = \|x_i - x_j\| = \sqrt{\sum_{l=1}^n (x_{i,l} - x_{j,l})^2} \quad (29)$$

Let, $\mathcal{W} = \{w_1, w_2, \dots, w_n\}$ be a transformation parameter of length n . Then, modified expression for weighted L^2 -distance is:

$$L_{\mathcal{W}}^2 = \sqrt{\sum_{l=1}^n w_l (x_{i,l} - x_{j,l})^2} \quad (30)$$

The weighted version of L^2 -distance has been widely used as a distance measure for classification and other algorithmic tasks [15, 44, 45, 51].

We show how to transform L^2 -distance's range directly by choosing a suitable transformation out of many possible transformations discussed in Sect. 3. This improves the model's performance. For validation purpose, we select k -means clustering [19] and employ it on LIBS amino acid data Sect. 6.3. Our choice of k -means clustering is to impact a very well known and widely used algorithm. We show how range transformation of L^2 -distance gives rise to improved k -means clustering performance by finding a suitable $\mathcal{W}$.

Open questions that we address concern: How is L^2 -distance range transformation related to k -means clustering? How does the range transformation impact k -means clustering? What kind of transformations can we choose? How to find a suitable transformation parameter $\mathcal{W}$ that gives the improved clustering results?

For n -many d -dimensional data points $\{x_i\}_{i=1}^n$, k -means clustering partitions the data into k clusters, where $k \leq n$. The objective is to minimize the sum of distances between each data point in the cluster and its cluster center (i.e., Sum of Squared Error (SSE)). This is expressed as an objective function:

$$\arg \min_K \sum_{i=1}^k \sum_{x \in K_i} \|x - \mu_i\|^2 \quad (31)$$

where K is the set of all clusters, and μ_i is the mean of data points belonging to cluster K_i .

Fig. 11 k -means clustering algorithm

The k -means algorithm produces exactly k different clusters maintaining maximum possible separation between them. While k is not known a priori and the prior probability for all k clusters is the same, an appropriate value for k can be found through validation. Therefore, k -means tries to produce clusters with approximately equal numbers of data points. Moreover, for k -means, the distribution of each variable is assumed to have the same variance (i.e., k -means tries to produce clusters which are spherical in shape). The algorithm for k -means clustering is given in Fig. 11.

5.1 Understanding L^2 -distance transformation in k -means

Consider data with two classes for k -means clustering as shown in Fig. 12. If there are subtle differences in classes, and cluster centers are unfavorably initialized as shown in Fig. 12a, k -means will yield poor clustering results by producing clusters containing both classes. An isotropic transformation of L^2 -distance matrix will transform all distances with the same scale factor, thus preserving the proportion of all relative distances (Fig. 12b, c). Therefore, in such scenario, an isotropic transformation cannot improve clustering results. Alternately, anisotropic transformation can put data points near to the right centers by shifting and orienting all L^2 -distances with different scale factors (Fig. 12d). This

Fig. 12 **a** Considering a scenario of 2 class data problem in k -means clustering. Here L^2 -distance of red point B from center $C1$ is bigger than its distance from center $C2$. The same is true for blue point A which is closer to center $C1$ than to center $C2$ **b** Isotropic dilation of L^2 -distances, shown on horizontal axis, does not help **c** Isotropic dilation of L^2 -distances, shown on vertical axis, does not help **d** anisotropic dilation of L^2 -distances can allow to put transformation at the right place and the right orientation so that the data points are closer to the right centers (color figure online)

Fig. 13 Some possible scenarios of bad clustering results by k -means clustering. **a** Clusters touching **b** clusters overlapping

refers back to our discussion on isotropic and anisotropic transformations in Sect. 3.4.

For good clustering, data points should be close to their respective centers and clusters should be far from each other. Depending on the data, k -means clustering gives poor results in the following scenarios:

- (i) Clusters are touching each other at boundaries in Fig. 13a.
It depends on the distributions of data points and assigned cluster centers. As k -means tries to make spherical clusters of equal size, it may create clusters which touch at each other's boundaries.
- (ii) Overlapping clusters in Fig. 13b.
It may happen if data points from different classes are equidistant from the cluster centers. As k -means tries to make clusters with roughly equal number of data points, it may yield overlapping clusters.
- (iii) Data points are assigned to wrong clusters.
If two data points from different classes have subtle differences they may be assigned to wrong cluster centers.

L^2 -distance range transformation may help in the following scenario:

- In the scenario when cluster boundaries are touching each other, a suitable choice of anisotropic transformation Sect. 3, on L^2 -distance matrix causes distances between data points and their respective centers to become smaller or bigger. Smaller L^2 -distances give compact clusters. Applied anisotropic transformation reconfigures L^2 -distance matrix such that the resulted clusters are separated from each other (Fig. 14). Transformed range of L^2 -distances rescales original L^2 -distances in a way so that the clusters contain data points which fall within that transformed range. This results in separated (compact) cluster (Fig. 14a). Other possibility of getting well-separated clusters is to reorientate L^2 -distances by anisotropic transformation without making clusters compact (Fig. 14b).
- In the scenario when clusters are overlapping, L^2 transformation may give separable compact clusters, nevertheless,

Fig. 14 Possible scenarios of separated clusters obtained by L^2 -distance range transformation in k -means clustering. **a** Compact clusters, **b** separated clusters

data points may not necessarily be assigned to the right cluster center as they are equidistant from centers. There is a 50% chances for an equidistant data point to be assigned to the correct center.

- If there are subtle differences in data distributions, this strategy is not successful.

5.2 Finding the transformation parameter by optimization

As we discussed in Sect. 5.1, our L^2 transformation approach can help for scenario when clusters are touching (Fig. 13a). A clustering is good if data points are close to their centroids and clusters are well separated from each other. This improvement is measured by computing Davies–Bouldin Index (DBI) which is defined as a function of the ratio comparing the sum of within cluster scatter and the between cluster separation [11].

Definition 3 (*Davies–Bouldin Index*) Let, $\bar{d}_i$ and $\bar{d}_j$ be the average distances between each point and the centroids in the i th and j th clusters, respectively, and let $d_{i,j}$ be the distance between the centroids of the i th and j th clusters. Then,

$$\text{DBI} = \frac{1}{k} \sum_{i=1}^k \max_{j \neq i} \left\{ \frac{\bar{d}_i + \bar{d}_j}{d_{i,j}} \right\}$$

where k is the number of clusters.

We employ an optimization approach to find a suitable transformation parameter, $\mathcal{W}$, responsible for an improvement in clustering results by transforming original L^2 .

Define optimization problem for objective function DBI w.r.t. $\mathcal{W}$. Let, $\mathcal{W} \in \mathbb{R}$ be a real vector with $n \geq 1$ and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be an objective function. Then, the unconstrained optimization problem is:

$$\min_{\mathcal{W}} f(X, \mathcal{W})$$

Where X is the data set.

Our optimization approach finds a suitable parameter $\mathcal{W}$ using anisotropic transformation of L^2 -distance matrix. Therefore, the transformed $L^2_{\mathcal{W}}$ -distances describe a model that improves k -means clustering results. This refers back to our strategy for finding a suitable model by transforming the range in backward direction (Fig. 2). The benefit of range transformation as an alternative to data transformation is that there is no early commitment about the data's geometry. Unlike data transformation, the implementation of range transformation focuses on altering the measurement tool, in our case, L^2 -distances. This is the primary contribution of this work.

We choose DBI for the interpretation and validation of consistency within clusters of data. It is important to note that other clustering validation methods may be employed (e.g., *Calinski-Harabasz criterion* also known as *variance ratio criterion* (VRC) [6], *Gap Value* [48], *Silhouette Value* [22,35]).

6 Experiments

As described in Sect. 1, we study range transformation approach in two ways: first changing the model to transform its range (forward direction) and then transform the range to change the model (backward direction) (Fig. 2).

For the *forward direction*, we employed synthetic data in our experiments Sect. 6.1. A goal of our experimentation was to demonstrate the impact of *dilation* afforded by CJSDs. We also set out to understand whether the dilation parameter's empirical behavior coincides with the theory by controlling for subtlety in the input data's distribution. We computed all CJSDs from input distributions from scratch and compared the results to our proofs. In the following discussion, we refer to Jensen–Shannon divergence measure as JSD, Chisini–Jensen–Shannon divergence measure as *CJSD*, the kernel function employing CJSD as *CJSD kernel*, and kernel density estimation as *KDE*.

For the *backward direction*, we choose a spectral data from the analysis of chemical compounds Sect. 6.3. A goal of our experimentation was to demonstrate the impact of *anisotropic dilation* on L^2 -norm in k -means clustering as described in Sect. 5.

6.1 Synthetic data generation

In the ideal case, classes in feature space consist of point clouds that are both compact and easily separable from the other classes. As noise increases, it becomes increasingly difficult to distinguish data points between classes, particularly those closer to the boundaries between two modes. We model noise as an increase in dispersion causing a high degree of overlap between the resulting point clouds. To model

Fig. 15 synthetic data for $\gamma = 0.1, 0.5, 0.9$

Fig. 16 Example of class overlap representing different subtlety problem

subtlety, we constructed a series of one-vs-one decision problems by combining multimodal point clouds consisting of classes whose dispersions differ in regular increments.

Define mixture distribution,

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^M w_i g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$$

where $\mathbf{x}$ is a d -dimensional continuous valued data vector, $w_i, i = 1, \dots, M$ are mixture weights and $g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), i = 1, \dots, M$ are component multivariate Gaussian densities in d -dimensions.

$$g(\mathbf{x}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_i|}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right\}$$

Beginning with the interval $[-1000, 1000] \in \mathbb{R}$, we set the means of our multimodal distribution by equally spacing apart two points per axis along 10-dimensions. This results in a total of 1024 modes in 10-dimensions. The Gaussian mixture is defined by setting the mean vectors to each mode's locus. The component distributions share a single isotropic covariance matrix $\boldsymbol{\Sigma}_i = \gamma \sigma^2 \mathbf{I}$ where σ^2 is initialized to the square of half the distance between the centers of adjacent modes. Setting γ along a schedule, 0.0, 0.1, $\dots$, 0.9, of 10% increments, we sample data sets of increasing smoothness.

We depict an example sampling distribution in Fig. 15 consisting of 4-modal 2-dimensional versions of the data with γ set to 0.1, 0.5, and 0.9. It is important to note in Fig. 15 that the vertical axis is the probability, depicted as a visual-

Fig. 17 High-intensity peaks of D-Serine (red) at certain wavelength in nanometer (nm) regions due to micro-well influence. Intensities in arbitrary unit (a.u.) of D-Serine (red) and Water (black) seem to be on top of other compounds due to the subtle differences in data classes. **a** LIBS experimental setup, **b** wavelength versus intensity plot (color figure online)

ization, of the data points in the (x_1, x_2) plane. As can be seen, an increase in γ results in overlap between modes due to variability. For example, a one-vs-one decision problem constructed by pairing the 4-modal $\gamma = 0.1$ data (red points) with the 4-modal $\gamma = 0.5$ data (green points) has the difference between the classes as more subtle than the decision problem constructed by pairing $\gamma = 0.1$ (red points) with $\gamma = 0.9$ (green points) (Fig. 16). Moreover, subtlety between pair of classes is not the same even though their difference in smoothnesses is the same (Fig. 16). For example, while the pairings $(\gamma = 0.1, \gamma = 0.2)$ and $(\gamma = 0.2, \gamma = 0.3)$ both have a difference in smoothness $\Delta_\gamma = 0.1$, the latter involves classes whose point clouds are more disperse and thus more noisy (stochasticity). In our synthetic data sets, we control for both subtlety between the classes as well as stochasticity. For our synthetic data, γ is the *smoothness* parameter. Moreover, for a one-vs-one decision problem, the difference in smoothness Δ_γ is the relative difference in smoothness between the paired data sets. Δ_γ measures the decision problem's subtlety.

Fig. 18 ISOMAP and LLE scatter plots with neighborhood size 15 on LIBS amino acid data. **a** ISOMAP, **b** LLE

Each test data set is a one-vs-one decision problem consisting of 2048 instances in 10-dimensions with equal priors over the classes. Given the unique pairings covering $\Delta_\gamma = 0.1, \dots, 0.9$ we generated a total of 45 tests.

6.2 CJSD kernels for SVM classification

We implemented *amplified* and *scaled* versions of CJSD kernels [41].

$$K_{\text{amplified}} = \text{CJSD}(P||Q)e^{-\frac{|x_i - x_j|^2}{2\sigma^2}} \quad (32)$$

$$K_{\text{scaled}} = e^{-\text{CJSD}(P||Q)\frac{|x_i - x_j|^2}{2\sigma^2}} \quad (33)$$

Because they modulate the radial basis function with information about the data's distribution, we performed SVM classification using WEKA's [18] sequential minimal optimization (SMO) algorithm [31] augmented with our custom kernels with RBF kernel as a comparative measure for CJSD's impact on classification performance. Where RBF takes pointwise information over the data set, augmentation with JSD considers the shape of the distribution over all data points while CJSD dilates the way in which the shape is considered. Moreover, given a complex decision problem with unknown nonlinearity, because the classification model must be able to shatter the data set, RBF kernel is a reasonable choice for SVM classification as it offers infinite VC dimen-

Fig. 19 (Top) Δ_γ versus CJSDs. (Green JSD_{AM} , Blue JSD_{GM} , Orange JSD_{HM}). (Middle) Δ_γ versus κ . (Bottom) Δ_γ versus Slope of κ . (Notice white line overlapping green line in middle and bottom subfigures confirming (18), (19), (20)) (color figure online)

sion. The RBF also carries the added benefit of constructing smooth solutions [5].

In all experiments, we performed tenfold cross-validation and report the sample mean and standard error. We estimated probability densities in our CJSDs using nonparametric KDE [29, 34, 43] where the bandwidth parameter is initialized with a diagonal covariance matrix computed from the data and grown, using Newton's method, to maximize the sum of leave-one-out log density [32]. It is important to note that data sets used in experiments for each tested CJSD kernel (AM, GM, and HM) are randomized with unique random number seeds. For fair comparison, the benchmark RBF experiment is run using the same randomized data sets.

6.3 LIBS data

Classification and identification of amino acids in aqueous solutions is important in the study of bio-macromolecules. Laser- Induced Breakdown Spectroscopy (LIBS) is a powerful analytical technique that provides information about the elemental composition of a given sample [17, 18]. The method uses an intense short laser pulse to break down the matrix of the target and to create a short-lived micro-plasma. During the cooling of plasma, atomic, ionic, and occasionally molecular constituents emit the spectra which is further collected and analyzed, allowing characterization of the sample content in Fig. 17a.

The distinction between amino acids from the perspective of their elemental composition is subtle as they appear very similar when measured by the emissions of their principal constituents, namely, hydrogen, carbon, nitrogen, and oxygen. In this work we analyze six compounds, namely, water, polysaccharide (Ficoll), aspartic acid (ASP), glutamic acid (Glu), cysteine (Cys), and D-Serine.

ASP, Cys, and Glu each consisted of 100 instances, D-Serine and water each consisted of 50 instances, and water consisted 270 instances. Each instance represent a single LIBS spectrum, produced from 13 different samples drawn from the six aforementioned compounds. Captured wavelengths remain consistent across all LIBS spectra. This resulted in 26,100 different wavelengths. In our experiment, we used the spectral intensity at each wavelength as a feature. Therefore, our data instances were vector valued in 26,100 dimensions.

We produced two different versions of data by employing ISOMAP [47] and LLE [36] with manifold neighborhood size 15. We computed ISOMAP scree plot which we used to select the top 3 principal components that explain over 98.5% variance. We choose top 3 principal components for both ISOMAP and LLE. From our 3D scatter plots in Fig. 18, it is obvious that these reduced data sets are not the same. Therefore, our experimental results with k -means clustering are tested on two different data sets.

7 Results

We set out to understand the performance improvement afforded by CJSDs. After deriving dilation parameter and the relationship between CJSDs, it was important to examine how the theory plays out empirically. We report results for CJSD output versus Δ_γ , Δ_γ versus κ , and Δ_γ versus Δ_κ (Fig. 19 top, middle, and bottom, respectively). Directing attention to results for CJSD versus Δ_γ , upper left subplot, the difference between CJSD outputs remains constant. This confirms Proposition 3 because the relative relationship between AM, GM, and HM versions is maintained regardless

of Δ_γ . Moreover, as variability in the data increases (reading top to bottom, left to right) Proposition 3 still holds. This means Proposition 3 is maintained as variability (or stochasticity) in the data increases. Results for Δ_γ versus κ show that the order of dilations coincides empirically with results derived in (18), (19), and (20) and holds regardless of difference in subtlety Δ_γ . That is, the difference in scaling κ between CJSD versions GM–AM and HM–GM is half the difference in scaling between HM and AM. Finally, in results for Δ_γ versus Δ_κ , as we increase subtlety, going to the right, the difference in value Δ_κ for the CJSD AM, GM, and HM versions preserves the aforementioned 2:1 ratio.

SVM classification results appear in Tables 1 and 2. As can be seen, in all cases, CJSDs outperform SVM with RBF Kernel with statistical significance at the 95% confidence level. Out of generated 45 tests with unique pairing of smoothness $\Delta_\gamma = 0.1, 0.2, \dots, 0.9$, for roughly 65% experiments, scaled and/or amplified versions of CJSD kernels for GM and HM give statistically significantly better performance than amplified version of CJSD kernel for AM. For the remaining experiments, both amplified and scaled versions of CJSD kernels give similar performance for AM, GM and HM.

7.1 Distortion in L^2 -distance

We discussed the role of L^2 -distance in k -means clustering. Now, in order to understand the role of L^2 -distance transformation in k -means clustering, recall transformation parameter $\mathcal{W}$ described in Sect. 5. The k -means algorithms iterates and computes L^2 -distances between data points and cluster centers until it converges in Fig. 11. These L^2 -distances are represented in a matrix of size $n \times k$, where n is the number of data points and k is the number of clusters.

Define distortion as a change in L^2 -distances due to introduced range transformation parameter $\mathcal{W}$. Let distortion $\Delta L = L_{\mathcal{W}}^2 - L^2$, where $L_{\mathcal{W}}^2$ is the transformed L^2 defined in (30). We compute distortions, for ISOMAP & LLE data sets, and plot them w.r.t. the number of iterations taken by k -means clustering to converge. This is done for clusters $k = 2:10$. The number of iteration steps taken by k -means can be different when we use L^2 and $L_{\mathcal{W}}^2$ distances. Generally, $L_{\mathcal{W}}^2$ version of k -means converges in fewer iterations than that of L^2 . We report distortion ΔL for the number of k -means iterations needed for the $L_{\mathcal{W}}^2$ version to converge. This was done to measure the impact of anisotropic transformation using comparative performance between L^2 and $L_{\mathcal{W}}^2$ versions of k -means. Therefore, we consider at most 10 iterations for computing distortion ΔL . We use an optimization approach in order to find $L_{\mathcal{W}}^2$.

From the subplots, we notice that ΔL decreases as the number of iterations increases and becomes stable after a few iterations (Fig. 20). This indicates that our optimization algorithm, defined in Sect. 5.2, converges toward an optimal

Table 1 CJSJ kernels SVM classification accuracies

Datasets	Kernel	A.M.(%)	G.M.(%)	H.M.(%)
D000 vs. D010	Amplify	59.6 ± 0.9	59.5 ± 0.4	60.2 ± 1.1
	Scaled	57.8 ± 0.9	82.7 ± 0.4	57 ± 0.7
D000 vs. D020	Amplify	68.3 ± 0.7	68 ± 1.4	67.3 ± 1.0
	Scaled	68.6 ± 0.9	83.2 ± 0.8	83.3 ± 0.8
D000 vs. D030	Amplify	78.1 ± 1.0	81.6 ± 0.9	81.9 ± 0.4
	Scaled	81.6 ± 0.9	83.7 ± 0.7	83.8 ± 0.9
D000 vs. D040	Amplify	88.7 ± 0.9	87.5 ± 0.8	87.8 ± 1.0
	Scaled	88.8 ± 0.7	83.1 ± 0.8	83.5 ± 0.6
D000 vs. D050	Amplify	92.5 ± 0.5	92.3 ± 0.5	92.4 ± 0.7
	Scaled	91.8 ± 0.9	86.9 ± 1.3	83.7 ± 0.7
D000 vs. D060	Amplify	95.7 ± 0.4	95.8 ± 0.6	95.7 ± 0.4
	Scaled	95.7 ± 0.4	93.6 ± 1.8	86.6 ± 1.9
D000 vs. D070	Amplify	97.3 ± 0.2	97.4 ± 0.4	97.5 ± 0.3
	Scaled	95.7 ± 0.2	97.4 ± 0.4	96.5 ± 1.1
D000 vs. D080	Amplify	98 ± 0.3	98 ± 0.3	98 ± 0.3
	Scaled	98 ± 0.3	98.1 ± 0.3	98.2 ± 0.3
D000 vs. D090	Amplify	98.9 ± 0.2	98.9 ± 0.2	98.9 ± 0.3
	Scaled	98.9 ± 0.2	98.9 ± 0.2	99.1 ± 0.3
D010 vs. D020	Amplify	63.2 ± 1.3	95.6 ± 0.8	63.8 ± 1.3
	Scaled	63.2 ± 1.3	69.3 ± 1.1	68.7 ± 1.0
D010 vs. D030	Amplify	72.5 ± 1.3	77.1 ± 0.9	78.1 ± 0.8
	Scaled	74.7 ± 1.1	79.1 ± 0.8	75.7 ± 0.7
D010 vs. D040	Amplify	82.8 ± 1.2	84.7 ± 0.9	85.6 ± 0.7
	Scaled	82.5 ± 0.9	80 ± 0.9	74.4 ± 1.2
D010 vs. D050	Amplify	89.2 ± 0.6	89.5 ± 0.7	90.6 ± 0.5
	Scaled	87.8 ± 0.7	81 ± 1	75.5 ± 0.7
D010 vs. D060	Amplify	94.5 ± 0.3	93.6 ± 0.7	93.3 ± 0.3
	Scaled	93.3 ± 0.4	80.4 ± 0.9	73.1 ± 1.2
D010 vs. D070	Amplify	96.6 ± 0.4	96.2 ± 0.4	96.2 ± 0.4
	Scaled	95.7 ± 0.6	80.6 ± 0.9	74.2 ± 1.1
D010 vs. D080	Amplify	97.5 ± 0.4	96.9 ± 0.2	96.6 ± 0.4
	Scaled	96.6 ± 0.4	81.3 ± 0.5	72.9 ± 0.6
D010 vs. D090	Amplify	98.7 ± 0.2	98.1 ± 0.3	98 ± 0.23
	Scaled	97.9 ± 0.3	81.7 ± 0.9	75.7 ± 0.7
D020 vs. D030	Amplify	61.6 ± 0.6	63.8 ± 1.0	63.3 ± 1.2
	Scaled	61.4 ± 1.1	64.7 ± 0.8	64.1 ± 1.4
D020 vs. D040	Amplify	70 ± 1.0	74.3 ± 0.6	74.5 ± 0.6
	Scaled	72.8 ± 0.6	74.2 ± 0.6	68.2 ± 0.8
D020 vs. D050	Amplify	76.8 ± 1.0	80.9 ± 0.6	80.5 ± 0.9
	Scaled	80 ± 0.7	80.5 ± 0.8	75.7 ± 0.7
D020 vs. D060	Amplify	86.9 ± 0.7	88.8 ± 0.9	88.2 ± 0.9
	Scaled	87.5 ± 0.8	87.4 ± 1.2	76.8 ± 1.7
D020 vs. D070	Amplify	91.4 ± 0.6	91.6 ± 0.6	91.6 ± 0.7
	Scaled	91.3 ± 0.7	90.7 ± 0.7	81.1 ± 2.2
D020 vs. D080	Amplify	94.5 ± 0.6	93.9 ± 0.5	94.2 ± 0.6
	Scaled	93.7 ± 0.6	90.9 ± 0.8	83.7 ± 1.3
D020 vs. D090	Amplify	96.5 ± 0.5	95.8 ± 0.5	95.3 ± 0.4
	Scaled	95.7 ± 0.5	94.6 ± 0.5	87.2 ± 0.8
D030 vs. D040	Amplify	56.4 ± 1.0	59.2 ± 1.0	58.1 ± 1.1
	Scaled	57.8 ± 1.4	59.5 ± 0.9	57.7 ± 0.6
D030 vs. D050	Amplify	61.8 ± 1.1	67.1 ± 0.9	66.2 ± 0.8
	Scaled	67 ± 1.2	67.5 ± 0.7	67.5 ± 0.8
D030 vs. D060	Amplify	74.1 ± 0.8	77.4 ± 0.9	76.8 ± 0.6
	Scaled	77.2 ± 0.8	77.2 ± 0.9	75.5 ± 0.9
D030 vs. D070	Amplify	81.3 ± 1.0	82.9 ± 0.7	83.1 ± 0.5
	Scaled	84 ± 0.7	82.7 ± 0.7	80.7 ± 0.6
D030 vs. D080	Amplify	86.4 ± 0.7	88.3 ± 0.6	87 ± 1.0
	Scaled	88.6 ± 0.7	87.8 ± 0.6	82.9 ± 1.2
D030 vs. D090	Amplify	90.1 ± 0.6	91.4 ± 0.8	90.6 ± 0.6
	Scaled	91.3 ± 0.4	89.5 ± 0.7	86.4 ± 0.9
D040 vs. D050	Amplify	54.2 ± 1.0	56.4 ± 1.0	56.4 ± 0.7
	Scaled	56.4 ± 1.1	56.6 ± 1.1	56.8 ± 0.9
D040 vs. D060	Amplify	62.4 ± 0.7	66.9 ± 1.0	67 ± 0.9
	Scaled	66.8 ± 0.8	66.2 ± 1.0	66.2 ± 1.0
D040 vs. D070	Amplify	69.8 ± 0.9	74.5 ± 1.0	74.2 ± 0.9
	Scaled	75.1 ± 0.6	74.1 ± 0.9	72.5 ± 1.0
D040 vs. D080	Amplify	76.4 ± 0.9	80.5 ± 1.0	80.5 ± 0.8
	Scaled	81.6 ± 0.8	80.1 ± 1.1	77.3 ± 0.7
D040 vs. D090	Amplify	81.2 ± 0.8	84.6 ± 0.4	84.4 ± 0.8
	Scaled	85.1 ± 0.9	84 ± 0.5	81.1 ± 0.7
D050 vs. D060	Amplify	54.6 ± 0.5	56.2 ± 0.7	57 ± 0.9
	Scaled	55.1 ± 1	57 ± 0.7	57.6 ± 0.7
D050 vs. D070	Amplify	60.6 ± 1.1	66 ± 0.9	64.2 ± 0.6
	Scaled	66.2 ± 1.1	65.9 ± 1.0	64.4 ± 0.7
D050 vs. D080	Amplify	67 ± 0.7	72.8 ± 1.2	71.4 ± 0.8
	Scaled	74.9 ± 0.7	73.2 ± 1.2	70.1 ± 0.9
D050 vs. D090	Amplify	72.9 ± 0.8	78.3 ± 0.8	77.8 ± 0.6
	Scaled	79 ± 0.8	78.3 ± 0.9	75.1 ± 0.7
D060 vs. D070	Amplify	53.6 ± 0.8	56.4 ± 0.9	56.3 ± 0.9
	Scaled	57.9 ± 1	56.4 ± 0.7	56.4 ± 0.7
D060 vs. D080	Amplify	59.7 ± 0.8	63.6 ± 0.7	62.1 ± 0.8
	Scaled	64.8 ± 0.9	63.8 ± 0.6	61.2 ± 1.2
D060 vs. D090	Amplify	65.8 ± 1.2	69.9 ± 0.9	68.9 ± 0.9
	Scaled	72 ± 1.4	69.8 ± 0.9	68 ± 0.9
D070 vs. D080	Amplify	54.6 ± 1.4	56.3 ± 0.6	55.5 ± 1.2
	Scaled	56.5 ± 1.1	56.3 ± 0.9	55.2 ± 1.1
D070 vs. D090	Amplify	57.3 ± 0.8	62.2 ± 0.8	62.2 ± 1.1
	Scaled	61.9 ± 1	62.5 ± 0.9	61.9 ± 1.1
D080 vs. D090	Amplify	53.1 ± 1.0	52.3 ± 0.8	52.9 ± 0.8
	Scaled	53.9 ± 0.7	50.6 ± 0.6	53.2 ± 1.1

Table 2 RBF kernels SVM classification accuracies

Datasets	Kernel	A.M.(%)	G.M.(%)	H.M.(%)
D000 vs. D010	Amplify	59.6 ± 0.9	59.5 ± 0.4	60.2 ± 1.1
	Scaled	57.8 ± 0.9	82.7 ± 0.4	57 ± 0.7
D000 vs. D020	Amplify	68.3 ± 0.7	68 ± 1.4	67.3 ± 1.0
	Scaled	68.6 ± 0.9	83.2 ± 0.8	83.3 ± 0.8
D000 vs. D030	Amplify	78.1 ± 1.0	81.6 ± 0.9	81.9 ± 0.4
	Scaled	81.6 ± 0.9	83.7 ± 0.7	83.8 ± 0.9
D000 vs. D040	Amplify	88.7 ± 0.9	87.5 ± 0.8	87.8 ± 1.0
	Scaled	88.8 ± 0.7	83.1 ± 0.8	83.5 ± 0.6
D000 vs. D050	Amplify	92.5 ± 0.5	92.3 ± 0.5	92.4 ± 0.7
	Scaled	91.8 ± 0.9	86.9 ± 1.3	83.7 ± 0.7
D000 vs. D060	Amplify	95.7 ± 0.4	95.8 ± 0.6	95.7 ± 0.4
	Scaled	95.7 ± 0.4	93.6 ± 1.8	86.6 ± 1.9
D000 vs. D070	Amplify	97.3 ± 0.2	97.4 ± 0.4	97.5 ± 0.3
	Scaled	95.7 ± 0.2	97.4 ± 0.4	96.5 ± 1.1
D000 vs. D080	Amplify	98 ± 0.3	98 ± 0.3	98 ± 0.3
	Scaled	98 ± 0.3	98.1 ± 0.3	98.2 ± 0.3
D000 vs. D090	Amplify	98.9 ± 0.2	98.9 ± 0.2	98.9 ± 0.3
	Scaled	98.9 ± 0.2	98.9 ± 0.2	99.1 ± 0.3
D010 vs. D020	Amplify	63.2 ± 1.3	95.6 ± 0.8	63.8 ± 1.3
	Scaled	63.2 ± 1.3	69.3 ± 1.1	68.7 ± 1.0
D010 vs. D030	Amplify	72.5 ± 1.3	77.1 ± 0.9	78.1 ± 0.8
	Scaled	74.7 ± 1.1	79.1 ± 0.8	75.7 ± 0.7
D010 vs. D040	Amplify	82.8 ± 1.2	84.7 ± 0.9	85.6 ± 0.7
	Scaled	82.5 ± 0.9	80 ± 0.9	74.4 ± 1.2
D010 vs. D050	Amplify	89.2 ± 0.6	89.5 ± 0.7	90.6 ± 0.5
	Scaled	87.8 ± 0.7	81 ± 1	75.5 ± 0.7
D010 vs. D060	Amplify	94.5 ± 0.3	93.6 ± 0.7	93.3 ± 0.3
	Scaled	93.3 ± 0.4	80.4 ± 0.9	73.1 ± 1.2
D010 vs. D070	Amplify	96.6 ± 0.4	96.2 ± 0.4	96.2 ± 0.4
	Scaled	95.7 ± 0.6	80.6 ± 0.9	74.2 ± 1.1
D010 vs. D080	Amplify	97.5 ± 0.4	96.9 ± 0.2	96.6 ± 0.4
	Scaled	96.6 ± 0.4	81.3 ± 0.5	72.9 ± 0.6
D010 vs. D090	Amplify	98.7 ± 0.2	98.1 ± 0.3	98 ± 0.23
	Scaled	97.9 ± 0.3	81.7 ± 0.9	75.7 ± 0.7
D020 vs. D030	Amplify	61.6 ± 0.6	63.8 ± 1.0	63.3 ± 1.2
	Scaled	61.4 ± 1.1	64.7 ± 0.8	64.1 ± 1.4
D020 vs. D040	Amplify	70 ± 1.0	74.3 ± 0.6	74.5 ± 0.6
	Scaled	72.8 ± 0.6	74.2 ± 0.6	68.2 ± 0.8
D020 vs. D050	Amplify	76.8 ± 1.0	80.9 ± 0.6	80.5 ± 0.9
	Scaled	80 ± 0.7	80.5 ± 0.8	75.7 ± 0.7
D020 vs. D060	Amplify	86.9 ± 0.7	88.8 ± 0.9	88.2 ± 0.9
	Scaled	87.5 ± 0.8	87.4 ± 1.2	76.8 ± 1.7
D020 vs. D070	Amplify	91.4 ± 0.6	91.6 ± 0.6	91.6 ± 0.7
	Scaled	91.3 ± 0.7	90.7 ± 0.7	81.1 ± 2.2
D020 vs. D080	Amplify	94.5 ± 0.6	93.9 ± 0.5	94.2 ± 0.6
	Scaled	93.7 ± 0.6	90.9 ± 0.8	83.7 ± 1.3
D020 vs. D090	Amplify	96.5 ± 0.5	95.8 ± 0.5	95.3 ± 0.4
	Scaled	95.7 ± 0.5	94.6 ± 0.5	87.2 ± 0.8
D030 vs. D040	Amplify	56.4 ± 1.0	59.2 ± 1.0	58.1 ± 1.1
	Scaled	57.8 ± 1.4	59.5 ± 0.9	57.7 ± 0.6
D030 vs. D050	Amplify	61.8 ± 1.1	67.1 ± 0.9	66.2 ± 0.8
	Scaled	67 ± 1.2	67.5 ± 0.7	67.5 ± 0.8
D030 vs. D060	Amplify	74.1 ± 0.8	77.4 ± 0.9	76.8 ± 0.6
	Scaled	77.2 ± 0.8	77.2 ± 0.9	75.5 ± 0.9
D030 vs. D070	Amplify	81.3 ± 1.0	82.9 ± 0.7	83.1 ± 0.5
	Scaled	84 ± 0.7	82.7 ± 0.7	80.7 ± 0.6
D030 vs. D080	Amplify	86.4 ± 0.7	88.3 ± 0.6	87 ± 1.0
	Scaled	88.6 ± 0.7	87.8 ± 0.6	82.9 ± 1.2
D030 vs. D090	Amplify	90.1 ± 0.6	91.4 ± 0.8	90.6 ± 0.6
	Scaled	91.3 ± 0.4	89.5 ± 0.7	86.4 ± 0.9
D040 vs. D050	Amplify	54.2 ± 1.0	56.4 ± 1.0	56.4 ± 0.7
	Scaled	56.4 ± 1.1	56.6 ± 1.1	56.8 ± 0.9
D040 vs. D060	Amplify	62.4 ± 0.7	66.9 ± 1.0	67 ± 0.9
	Scaled	66.8 ± 0.8	66.2 ± 1.0	66.2 ± 1.0
D040 vs. D070	Amplify	69.8 ± 0.9	74.5 ± 1.0	74.2 ± 0.9
	Scaled	75.1 ± 0.6	74.1 ± 0.9	72.5 ± 1.0
D040 vs. D080	Amplify	76.4 ± 0.9	80.5 ± 1.0	80.5 ± 0.8
	Scaled	81.6 ± 0.8	80.1 ± 1.1	77.3 ± 0.7
D040 vs. D090	Amplify	81.2 ± 0.8	84.6 ± 0.4	84.4 ± 0.8
	Scaled	85.1 ± 0.9	84 ± 0.5	81.1 ± 0.7
D050 vs. D060	Amplify	54.6 ± 0.5	56.2 ± 0.7	57 ± 0.9
	Scaled	55.1 ± 1	57 ± 0.7	57.6 ± 0.7
D050 vs. D070	Amplify	60.6 ± 1.1	66 ± 0.9	64.2 ± 0.6
	Scaled	66.2 ± 1.1	65.9 ± 1.0	64.4 ± 0.7
D050 vs. D080	Amplify	67 ± 0.7	72.8 ± 1.2	71.4 ± 0.8
	Scaled	74.9 ± 0.7	73.2 ± 1.2	70.1 ± 0.9
D050 vs. D090	Amplify	72.9 ± 0.8	78.3 ± 0.8	77.8 ± 0.6
	Scaled	79 ± 0.8	78.3 ± 0.9	75.1 ± 0.7
D060 vs. D070	Amplify	53.6 ± 0.8	56.4 ± 0.9	56.3 ± 0.9
	Scaled	57.9 ± 1	56.4 ± 0.7	56.4 ± 0.7
D060 vs. D080	Amplify	59.7 ± 0.8	63.6 ± 0.7	62.1 ± 0.8
	Scaled	64.8 ± 0.9	63.8 ± 0.6	61.2 ± 1.2
D060 vs. D090	Amplify	65.8 ± 1.2	69.9 ± 0.9	68.9 ± 0.9
	Scaled	72 ± 1.4	69.8 ± 0.9	68 ± 0.9
D070 vs. D080	Amplify	54.6 ± 1.4	56.3 ± 0.6	55.5 ± 1.2
	Scaled	56.5 ± 1.1	56.3 ± 0.9	55.2 ± 1.1
D070 vs. D090	Amplify	57.3 ± 0.8	62.2 ± 0.8	62.2 ± 1.1
	Scaled	61.9 ± 1	62.5 ± 0.9	61.9 ± 1.1
D080 vs. D090	Amplify	53.1 ± 1.0	52.3 ± 0.8	52.9 ± 0.8
	Scaled	53.9 ± 0.7	50.6 ± 0.6	53.2 ± 1.1

Fig. 20 Applied distortion in L^2 -distance due to its range transformation reduces and comes to a Pareto after a few iterations. **a** ISOMAP **b** LLE

Fig. 21 DBI decreases and becomes stable (less fluctuations) with increasing distortion in L^2 -distance due to its range transformation. **a** ISOMAP, **b** LLE

Fig. 22 Pattern search, global search and multipoint search perform better than all tested optimization methods. Optimal DBI value (smaller) indicates better clustering than the case when no optimization (i.e., no L^2 transformation) was performed (red line). **a** Optimization results on ISOMAP data, **b** optimization results on LLE data (color figure online)

$\mathcal{W}$ required to introduce a necessary anisotropic transformation in L^2 -distances.

We analyzed the role of $L^2_{\mathcal{W}}$ in k -means clustering. Now, we analyze its impact on k -means clustering. In other words, we set out to understand that how clustering improves with the transformation of L^2 -distance. The improvement in clustering can be measured by computing Davies–Bouldin Index (DBI; smaller is better). We compute distortions ΔL for ISOMAP & LLE data sets and plot them w.r.t. DBI. This is done for clusters $k = 2:10$. Like before, we find optimal $\mathcal{W}$ which gives transformed L^2 -distances by using our optimization approach given in Sect. 5.2.

From the subplots, we notice that the ΔL versus DBI is not a smooth plot; DBI decreases non-monotonically (Fig. 21). This is an artifact of the optimization. However, as distortion increases, DBI values start becoming closer. This means that our optimization algorithm creates distortions in L^2 -distance until it finds a suitable $\mathcal{W}$ responsible for L^2 -distance transformation. This transformed $L^2_{\mathcal{W}}$ gives an optimal DBI.

7.2 Optimal DBI

Using our approach in Sect. 5.2, we employed many optimization algorithms to find a suitable method which gives

an optimal DBI value for some transformation parameter $\mathcal{W}$. Optimization methods can be characterized on the basis of search algorithm they employ. For example, linear and nonlinear, gradient (direct search) and non-gradient based, deterministic and stochastic start point, and constrained and unconstrained [12]. We compute optimal DBIs w.r.t. cluster numbers for both, ISOMAP & LLE data sets, and benchmark our DBI results against when no L^2 transformation was used. This is done for clusters $k = 2:10$. From Fig. 22 (best viewed in color), we notice that Pattern Search (blue line), Global Search (black line) and Multipoint Search (black asterisk line) give better clustering results (smaller DBI) versus when no optimization was used (red line, when no L^2 transformation was used for k -means clustering). While Pattern Search and Multipoint Search outperform other employed optimization methods for most k , Global Search outperforms for every k .

8 Conclusion

This work contributed a method for modifying a broad class of models to suit novel data through transformation of the range. We investigated different classes of transformations and evaluated their impact on machine learning tasks (classification and clustering). We demonstrated that by focusing on transformations of function's range, we can find a generally applicable way to change the function's properties to best suit the data.

Our contribution to machine learning is an approach to selecting suitable models by focusing on transformations of their range. In particular when faced with novel data with little knowledge about pattern phenomena (e.g., quantum effects due to photon energy absorption) and the needed resolution required between model outputs in order to discriminate between classes, a more general approach informed by resolution can be valuable. Given a certain degree of subtlety between classes within the data or between members of a cluster, one begins with the requisite resolution in distance measure needed for discriminability and then selects a transformation achieving it. By this approach a more broad class of functions is available for modeling. A new framework employing range transformations to improve models was developed.

We studied range transformation approach in two ways: first changing the model to transform its range and then to transform the range to change the model (Fig. 2).

Going in *forward direction*, because of their broad applicability in describing regularities in data, we focused on the development of range transformations arising from function edits of JSDs using Chisini means. The family of modified Jensen–Shannon divergences, or Chisini–Jensen–Shannon divergences (CJSDs), carry important properties responsible

for performance improvements. We addressed open questions related to the transformation of CJSD's range, properties and limitations of JSDs reconfigurations, data domains to which these reconfigurations are applicable in general, precise rescaling done by CJSDs between input distributions, the magnitude of this rescaling, and computing CJSDs implicitly.

Beginning with an initial focus on how reformulation of JSDs with Chisini mean rescaled their range, we proved a number of mathematical propositions that establish the governing parameter, namely the *dilation parameter* and *dilation mapping*. For empirical validation, we generated synthetic data sets controlling for unit step increments in variability and subtlety between the classes. We validated the dilation parameter and showed that it holds regardless of subtlety (something we control for in our synthetic data experiments). We control for subtlety using covariance in multimodal multivariate Gaussian). Our findings demonstrate utility of Chisini means in classification of complex data whose properties include stochasticity and subtlety. These findings explain the $AM < GM < HM$ ordering among CJSDs and sheds light on how the JSD distance metric spaces apart its outputs resulting in improvement in kernel performance. When the differences between classes in data are subtle, an understanding of dilation provides tools to begin suggesting the appropriate measure for drawing distinctions. For example, if subtlety is small, our results recommend using HM version of CJSD. Most importantly, our experiments confirm our theories concerning governing parameters as well as their utility as kernels in SVM classification. As the current work is theoretical in nature, future work will include evaluation of performance on real-world data sets.

After empirically validating our theoretical results in forward direction, we studied range transformation in *backward direction*. In doing so, we developed an optimization approach for finding a suitable model. With our approach, we showed that one can change the model by focusing on the transformation of the model's range. A validation using L^2 -distance range transformation was provided for k -means clustering on LIBS data. This work demonstrated the benefit of the range transformation as an alternative to data transformation in clustering methods. Our optimization approach finds a suitable parameter $\mathcal{W}$ using anisotropic transformation of L^2 -distance matrix. Therefore, the transformed $L^2_{\mathcal{W}}$ -distances describe a model that improves k -means clustering results. This refers back to our strategy for finding a suitable model by transforming the range in backward direction. The benefit of range transformation as an alternative to data transformation is that there is no early commitment about the data's geometry. Unlike data transformation, the implementation of range transformation focuses on altering the measurement tool, in our case, L^2 -distances. This is the primary contribution of this work.

L^2 -distance is widely applicable as a measuring tool. We demonstrated an anisotropic range transformation approach for this measure. Further investigation can be carried out for other domains beyond k -means clustering. Moreover, transformations for other measures, such as, Mahalanobis distance [27], Bhattacharyya distance (Hellinger distance) [2], and Manhattan distance [23] can be studied and analyzed for and beyond k -means clustering.

One distance measure can be reduced to other distance measure. For example, if in the definition of Mahalanobis distance, the covariance matrix is the identity matrix, it reduces to Euclidean distance. Future work will investigate variations in distance measures by using range transformation. In particular, we will study *what kind of transformations do these variations give?*

Finally, in this work we explored range transformation for the analysis of synthetic and spectral data sets, which carry properties of subtlety, stochasticity and unknown nonlinearities. Future work will explore range transformation for other data domains, such as, astronomical, video, and voice data each one with unique characteristics (noise, nonlinearity, nuances in data classes, etc.).

Acknowledgements The authors would like to thank Dr. Yuri Markushin, Dr. Poopalasingam Sivakumar, and Dr. Noureddine Melikechi for their assistance in understanding LIBS and LIBS experimental protocol.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

- Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. *J. Mach. Learn. Res.* **6**, 1705–1749 (2005)
- Bhattacharyya, A.: On a measure of divergence between two multinomial populations. *Sankhyā Indian J. Stat.* **401**–406 (1946)
- Blanc, C., Guitton, P., Schlick, C.: A methodology for description of geometrical deformations. In: *Proceedings of Pacific Graphics*, vol. 94 (1994)
- Brannon, R.: Kinematics: the mathematics of deformation. *Course Notes*, ME EN **6530** (2008)
- Buhmann, M.D.: Radial basis functions: theory and implementations. *Camb. Monogr. Appl. Comput. Math.* **12**, 147–165 (2004)
- Calinski, T., Harabasz, J.: A dendrite method for cluster analysis. *Commun. Stat. Theory Methods* **3**(1), 1–27 (1974)
- Chen, J., Thalmann, N.M., Tsang, Z., Thalmann, D.: *Fundamentals of computer graphics*. World Scientific, Singapore (1994)
- Chisini, O.: Sul concetto di media. *Periodico di Matematiche* **9**:2(4), 106–116 (1929)
- Cover, T.M., Thomas, J.A.: *Elements of Information Theory*. Wiley, New York (2012)
- Coxeter, H.S.M., Greitzer, S.L.: *Geometry Revisited*. MAA, Washington, DC (1967)
- Davies, D.L., Bouldin, D.W.: A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-1**(2), 224–227 (1979)
- Dixon, L.C.W., Szegő, G.P.: *Towards Global Optimisation*, vol. 2. North-Holland, Amsterdam (1978)
- Faloutsos, C., Lin, K.I.: Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. *SIGMOD Rec.* **24**(2), 163–174 (1995)
- Gelman, A., Hill, J.: *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, Cambridge (2006)
- Greenacre, M.J., Groenen, P.J.: Weighted Euclidean biplots. *J. Classif.* **33**(3), 442–459 (2016)
- Griffith, D.A.: Reformulating classical linear statistical models. In: *Advanced Spatial Statistics*, pp. 82–107. Springer Netherlands (1988)
- Güvenir, H.A., Altıngöve, S., Uysal, I., Erel, E.: Bankruptcy prediction using feature projection based classification. In: *Proceedings of SCI/ISAS* (1999)
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. *ACM SIGKDD Explor. Newsl.* **11**(1), 10–18 (2009)
- Hartigan, J.A., Wong, M.A.: Algorithm as 136: a k -means clustering algorithm. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **28**(1), 100–108 (1979)
- Inder, J.: New developments in generalized information measures. *Adv. Imag. Electron Phys.* **91**, 37–135 (1995)
- Kapur, J.N.: A comparative assessment of various measures of directed divergence. *Adv. Manag. Stud.* **3**(1), 1–16 (1984)
- Kaufman, L., Rousseeuw, P.J.: *Finding Groups in Data: An Introduction to Cluster Analysis*, vol. 344. Wiley, New York (2009)
- Krause, E.F.: *Taxicab Geometry: An Adventure in Non-Euclidean Geometry*. Courier Corporation, North Chelmsford (2012)
- Kullback, S.: *Information Theory and Statistics*, 2nd edn. Dover Publications, New York (1968)
- Kullback, S., Leibler, R.A.: On information and sufficiency. *Ann. Math. Stat.* **22**(1), 79–86 (1951)
- Lin, J.: Divergence measures based on the Shannon entropy. *IEEE Trans. Inf. Theory* **37**(1), 145–151 (1991)
- Mahalanobis, P.C.: On the generalised distance in statistics. *Proc. Natl. Inst. Sci. India* **2**(1), 49–55 (1936)
- Mortenson, M.: *Geometric Transformations for 3D Modeling*. Industrial Press Inc, Norwalk (2007)
- Parzen, E.: On estimation of a probability density function and mode. *Ann. Math. Stat.* **33**(3), 1065–1076 (1962)
- Pflug, G.C.: On distortion functionals. *Stat. Decis.* **24**(1/2006), 45–60 (2006)
- Platt, J.C.: Advances in kernel methods. In: Schölkopf, B., Burges, C.J.C., Smola, A.J. (eds.) *Advances in Kernel Methods*, chap. Fast Training of Support Vector Machines Using Sequential Minimal Optimization, pp. 185–208. MIT Press, Cambridge (1999)
- Raykar, V.C., Duraiswami, R.: Fast optimal bandwidth selection for kernel density estimation. In: *SDM*, pp. 524–528. SIAM (2006)
- Rodgers, J.L., Kohler, H.P.: Reformulating and simplifying the df analysis model. *Behav. Genet.* **35**(2), 211–217 (2005)
- Rosenblatt, M.: Remarks on some nonparametric estimates of a density function. *Ann. Math. Stat.* **27**(3), 832–837 (1956)
- Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987)
- Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *Science* **290**(5500), 2323–2326 (2000)
- Schlick, C.B.P.G.C.: A methodology for description of geometrical deformations. *Fundam. Comput. Graph.* **94** (1994)
- Shannon, C.: A mathematical theory of communication. *Bell Syst. Techn. J.* **27**(379–423), 623–656 (1948)
- Sharma, P., Holness, G.: Dilation of Chisini–Jensen–Shannon divergence. In: *3rd IEEE International Conference on Data Science and Advanced Analytics* (2016)

40. Sharma, P.K., Holness, G.: Dilation of Chisini–Jensen–Shannon divergences. In: 33rd International Conference on Machine Learning (ICML 2016) (2016)
41. Sharma, P.K., Holness, G., Markushin, Y., Melikechi, N.: A family of Chisini mean based Jensen–Shannon divergence kernels. In: 14th IEEE International Conference on Machine Learning and Applications. IEEE, Miami, FL (2015)
42. Simon, U.: Affine differential geometry. *Handb. Differ. Geom.* **1**, 905–961 (2000)
43. Simonoff, J.S.: *Smoothing Methods in Statistics*. Springer, Berlin (2012)
44. Singha, J., Das, K.: Indian sign language recognition using eigen value weighted Euclidean distance based classification technique. arXiv preprint [arXiv:1303.0634](https://arxiv.org/abs/1303.0634) (2013)
45. Smith, A.M.: A dual algorithm for the weighted Euclidean distance min–max location problem in R^2 and R^3 . ProQuest. Doctoral dissertation, Clemson University, South Carolina, USA (2009)
46. Taneja, I.J.: On a difference of Jensen inequality and its applications to mean divergence measures. [arXiv:math/0501302](https://arxiv.org/abs/math/0501302) (2005)
47. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. *Science* **290**(5500), 2319–2323 (2000)
48. Tibshirani, R., Walther, G., Hastie, T.: Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **63**(2), 411–423 (2001)
49. Vapnik, V.: *The Nature of Statistical Learning Theory*. Springer, Berlin (2013)
50. Wagstaff, A.: The demand for health: an empirical reformulation of the Grossman model. *Health Econ.* **2**(2), 189–198 (1993)
51. Wang, L., Zhang, Y., Feng, J.: On the Euclidean distance of images. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(8), 1334–1339 (2005)