Relaxed Exponential Kernels for Unsupervised Learning

Karim Abou-Moustafa[1], Mohak Shah[2]
, Fernando De La Torre[3], and Frank Ferrie[1]

[1] Centre of Intelligent Machines, McGill University,
3480 University street, Montréal, QC, H3A 2A7, CANADA.
{karimt,ferrie}@cim.mcgill.ca
[2] Accenture Technology Labs.
161 N. Clark street Chicago, IL, 60601, U.S.A.
mohak.shah@accenture.com
[3] The Robotics Institute, Carnegie Mellon Unviversity,
5000 Forbes Avenue, Pittsburg, PA 15213, U.S.A.
ftorre@cs.cmu.edu

Abstract. Many unsupervised learning algorithms make use of kernels that rely on the Euclidean distance between two samples. However, the Euclidean distance is optimal for Gaussian distributed data. In this paper, we relax the global Gaussian assumption made by the Euclidean distance, and propose a locale Gaussian modelling for the immediate neighbourhood of the samples, resulting in an augmented data space formed by the parameters of the local Gaussians. To this end, we propose a convolution kernel for the augmented data space. The factorisable nature of this kernel allows us to introduce (semi)-metrics for this space, which further derives relaxed versions of known kernels for this space. We present empirical results to validate the utility of the proposed localized approach in the context of spectral clustering. The key result of this paper is that this approach that combines the local Gaussian model with measures that adhere to metric properties, yields much better performance in different spectral clustering tasks.

1 Introduction

Many unsupervised learning algorithms rely on the exponential kernel K_E, and the Gaussian kernel K_G to measure the similarity between two input vectors[4] $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$. The Euclidean distance in K_E and K_G, however, has two implicit assumptions on the data under consideration. First, by expanding the squared norm $\|\mathbf{x}-\mathbf{y}\|^2$ to $(\mathbf{x}-\mathbf{y})^\top \mathbf{I}(\mathbf{x}-\mathbf{y})$, where $\mathbf{I}$ is the identity matrix, one directly obtains a special case of the generalized quadratic distance (GQD) $d(\mathbf{x}, \mathbf{y}; \mathbf{A}) =$

[4] **Notations:** Bold small letters $\mathbf{x}, \mathbf{y}$ are vectors. Bold capital letters $\mathbf{A}, \mathbf{B}$ are matrices. Calligraphic and double bold capital letters $\mathcal{X}$, $\mathcal{Y}$, $\mathbb{X}$, $\mathbb{Y}$ denote sets and/or spaces. Positive definite (PD) and positive semi-definite (PSD) matrices are denoted by $\mathbf{A} \succ 0$ and $\mathbf{A} \succeq 0$ respectively.

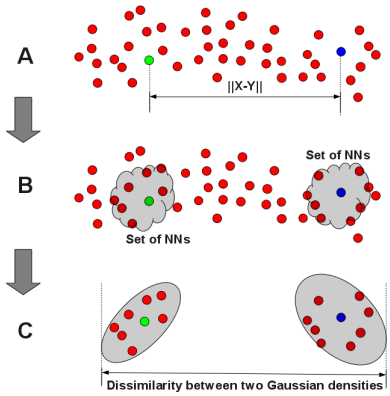

Fig. 1. (A) The exponential kernel K_E relies on the Euclidean distance between X (green) and Y (blue). (B) The local Gaussian assumption considers the few nearest neighbours (NNs) around X and Y, and then each set of NNs is modelled as a Gaussian distribution as in (C). The proposed relaxed kernels will rely on the dissimilarity (or difference) between the two Gaussian distributions instead of the Euclidean distance between X and Y.

$\sqrt{(\mathbf{x}-\mathbf{y})^\top \mathbf{A}(\mathbf{x}-\mathbf{y})}$, where $\mathbf{A}$ is a symmetric PD matrix. From a statistical vantage point, the Euclidean distance is the optimal metric if the data is generated from a spherical Gaussian distribution with unit variances – *the spherical assumption* – which is a hard to attain natural setting in real world data sets.

Second, the GQD has an inherent limitation for which the matrix $\mathbf{A}$ is constrained to be globally defined over the whole input space, which enforces the global Gaussian assumption of the data, or *the ellipsoidal assumption*. Besides that this constraint on $\mathbf{A}$ is restrictive, the ellipsoidal assumption is unjustified since a large Gaussian distribution with a full covariance matrix, does not yield a faithful modelling for the true empirical density of the data. In turn, this affects the relative distances between the samples, which finally affects the similarity evaluated by K_E and K_G.

In this paper, we propose to relax the constraint that enforces the global Gaussian assumption on the data. That is, as depicted in Figure (1), instead of being globally defined over all the data set, the Gaussian assumption is allowed to only hold in a local neighbourhood around each sample $\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^p$, where $\mathcal{X}$ is the input space. Note that the local Gaussian assumption, does not impose any constraints nor assumptions on the global data distribution. The local Gaussian assumption, however, associates with each $\mathbf{x}_i$ a symmetric PD matrix $\mathbf{A}_i$, which is the covariance matrix of the local Gaussian distribution centered at $\mathbf{x}_i$. In turn, this changes the structure of the data from the simple set of vectors $D = \{\mathbf{x}_i\}_{i=1}^n \subseteq \mathcal{X}$, to a new *augmented data set* $D_A = \{(\mathbf{x}_i, \mathbf{A}_i)\}_{i=1}^n \subseteq \mathbb{X}$ of the 2-tuples $(\mathbf{x}_i, \mathbf{A}_i)$. Note that all $\mathbf{A}_i$'s are defined in an unsupervised manner.

To this end, we propose a convolution kernel $K_{\mathbb{X}}$ [18] that measures the similarity between the inputs $(\mathbf{x}_i, \mathbf{A}_i)$ and $(\mathbf{x}_j, \mathbf{A}_j)$. The kernel $K_{\mathbb{X}}$ is an exponential

function of a dissimilarity measure for the 2-tuples $(\mathbf{x}_i, \mathbf{A}_i)$. Due to the factorizable nature of $K_{\mathbb{X}}$, it turns that $K_{\mathbb{X}}$ derives a set of metrics and semi-metrics on the augmented space $\mathbb{X}$, which further derive a set of relaxed kernels for $\mathbb{X}$. Interestingly, these (semi-)metrics are based on divergence measures of probability distributions, and the Riemannian metric for symmetric PD matrices [16]. Moreover, we show that using the exponential function in $K_{\mathbb{X}}$, the space $\mathbb{X}$ is isometrically embeddable into a Hilbert space $\mathcal{H}$ [17].

Preliminaries In order to make the paper self-contained, we find it necessary to introduce the following definitions. A metric space is an ordered pair $(\mathcal{M}, d)$ where $\mathcal{M}$ is a non-empty set, and d is a distance function, or a metric, defined as $d : \mathcal{M} \times \mathcal{M} \mapsto \mathbb{R}$, and $\forall\ a, b, c \in \mathcal{M}$, the following Axioms hold: (1) $d(a, b) \geq 0$, (2) $d(a, a) = 0$, (3) $d(a, b) = 0$ iff $a = b$, (4) symmetry $d(a, b) = d(b, a)$, and (5) the triangle inequality $d(a, c) \leq d(a, b) + d(b, c)$. A semi-metric distance satisfies Axioms (1), (2) and (4) only. That is, the triangle inequality need not hold for semi-metrics, and $d(a, b)$ can be zero for $a \neq b$. For instance, $\|\mathbf{x} - \mathbf{y}\|_2$ in K_E is a metric, but $\|\mathbf{x} - \mathbf{y}\|_2^2$ in K_G is a semi-metric. Similarly for the GQD, $d^2(\mathbf{x}, \mathbf{y}; \mathbf{A})$ is a semi-metric, and if $\mathbf{A}$ is not strictly PD, then $d(\mathbf{x}, \mathbf{y}; \mathbf{A})$ is also a semi-metric. Note that the definition of a metric space is independent from whether $\mathcal{M}$ is equipped with an inner product or not.

Axioms (1) & (2) produce the positive semi–definiteness (PSD) of d, and hence metrics and semi-metrics are both PSD. Note that this PSD property is only valid for metrics and semi-metrics due to their Axiomitic definition above, and can not be generalized to other PSD function as defined in the following.

A necessary and sufficient condition to guarantee that a symmetric similarity function K is a kernel function over $\mathcal{X}$, is that K should be PSD[5]. This ensures the existence of a mapping $\boldsymbol{\phi} : \mathcal{X} \mapsto \mathcal{H}$, where $\mathcal{H}$ is a Hilbert space called the feature space, in which K turns into an inner product: $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \boldsymbol{\phi}(\mathbf{x}_i), \boldsymbol{\phi}(\mathbf{x}_j) \rangle$.

The family of p–dimensional Gaussian distributions is denoted by $\mathbb{G}_p$, and for $\mathcal{G} \in \mathbb{G}_p$, it is defined as:

$$\mathcal{G}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\},$$

where $|\cdot|$ is the determinant, $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^p$, $\boldsymbol{\Sigma} \in \mathbb{S}_{++}^{p \times p}$, and $\mathbb{S}_{++}^{p \times p}$ is the manifold of symmetric PD matrices.

2 The local Gaussian assumption

Our proposal for relaxing the constraint on matrix $\mathbf{A}$ in the GQD is equivalent to relaxing the global Gaussian assumption on the data to be only valid in a small neighbourhood around each sample $\mathbf{x}_i \in \mathcal{X}$. Note that this mild assumption on the local distribution around each $\mathbf{x}_i$ does not impose any constraints nor assumptions on the global data distribution. To realize the local Gaussian assumption, each $\mathbf{x}_i$ is associated with a symmetric matrix $\mathbf{A}_i \succ 0$ defined as:

$$\mathbf{A}_i = \frac{1}{m-1} \sum_{\mathbf{x}^j \in \mathcal{N}_i}^{m} (\mathbf{x}^j - \mathbf{x}_i)(\mathbf{x}^j - \mathbf{x}_i)^\top + \gamma \mathbf{I}\ , \tag{1}$$

[5] For the set $\mathcal{X}$ and for any set of real numbers $a_1, \ldots, a_n$, the function K must satisfy the following: $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.

where $\mathbf{x}^j \in \mathcal{X}$, $\mathcal{N}_i = \{\mathbf{x}^j\}_{j=1}^m$ is the set of m nearest neighbours (NNs) to $\mathbf{x}_i$, and $0 < \gamma \in \mathbb{R}$ is a regularization parameter. The regularization here is necessary to avoid the expected rank deficiencies in $\mathbf{A}_i$'s, which are due to the small number of NNs considered around $\mathbf{x}_i$, together with the high dimensionality of the data[6], and hence, this helps avoid over-fitting and outlier reliance. The definition of $\mathbf{A}_i$ in (1) is simply the average variance–covariance matrix between $\mathbf{x}_i$ and its m NNs. Hence, the local Gaussian assumption, depicted in Figure (1), can be seen as anchoring a Gaussian density $\mathcal{G}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ at point $\mathbf{x}_i$, where its mean $\boldsymbol{\mu}_i \equiv \mathbf{x}_i$ and its covariance matrix $\boldsymbol{\Sigma}_i \equiv \mathbf{A}_i$. The local Gaussian assumption can be taken further and extended in the spirit of manifold Parzen windows [19] by including $\mathbf{x}_i$ in $\mathcal{N}_i$, and define $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ as follows:

$$\boldsymbol{\mu}_i \equiv \hat{\boldsymbol{\mu}}_i = \frac{1}{m+1} \sum_{\mathbf{x}^j \in \mathcal{N}_i} \mathbf{x}^j, \quad \text{and} \tag{2}$$

$$\boldsymbol{\Sigma}_i \equiv \hat{\boldsymbol{\Sigma}}_i = \frac{1}{m} \sum_{\mathbf{x}^j \in \mathcal{N}_i} (\mathbf{x}^j - \hat{\boldsymbol{\mu}}_i)(\mathbf{x}^j - \hat{\boldsymbol{\mu}}_i)^\top + \gamma \mathbf{I} \; . \tag{3}$$

This can be seen as a local smoothing for the data, combined with local feature extraction by means of a generative model, where the features are the parameters $\hat{\boldsymbol{\mu}}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ for each $\mathbf{x}_i \in \mathcal{X}$. Note that $\mathbf{A}_i$ and $\hat{\boldsymbol{\Sigma}}_i$ are defined in an unsupervised manner, however when auxiliary information is available in the form of labels or side information, they can be defined in a supervised or a semi–supervised manner.

The result of the local Gaussian assumption introduces a new component $\mathbf{A}_i$ for each $\mathbf{x}_i \in \mathcal{X}$ which changes the structure of the input data from the set of vectors $D = \{\mathbf{x}_i\}_{i=1}^n$ to an augmented data set $D_A = \{(\mathbf{x}_i, \mathbf{A}_i)\}_{i=1}^n \subseteq \mathbb{X}$ of 2-tuples $(\mathbf{x}_i, \mathbf{A}_i)$. This change in the data structure, in turn, requires a change in K_E and K_G which can only operate on the first element of the 2-tuples $(\mathbf{x}_i, \mathbf{A}_i)$ – elements in $\mathbb{R}^p$ – and not the symmetric matrix $\mathbf{A}_i \succ 0$.

Note that the augmented space $\mathbb{X}$ implicitly represents the parameters for the set of local Gaussians $\mathscr{G} = \{\mathcal{G}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{i=1}^n$, which will be referred to as *the dual perspective* for $\mathbb{X}$. In order to avoid any future confusion in the notations, this will be the default definition for $\mathbb{X}$, where implicitly, $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \equiv (\mathbf{x}_i, \mathbf{A}_i)$, or $(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \equiv (\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)$.

3 A convolution kernel for the space $\mathbb{X}$

The framework of convolution kernels suggests that a possible kernel for the space $\mathbb{X}$ can have the following structure [18]:

$$K_{\mathbb{X}}\{(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), (\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\} = K_\mu(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) K_\Sigma(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j),$$

where K_μ and K_Σ are symmetric PSD kernels, which yields that $K_{\mathbb{X}}$ is symmetric and PSD as well. Our approach for defining K_μ and K_Σ is based on the definition

[6] Note that γ is unique for all $\mathbf{A}_i$'s.

of K_E, which is an exponential function of the Euclidean distance between its two inputs. Due to the PSD and symmetry properties of (semi-)metrics (Axioms (1), (2), & (3)), it follows that K_E is symmetric and PSD. This result is due to Theorem (4) in [17] which states that:

Theorem 1. *The most general positive function $f(x)$ which is bounded away from zero and whose positive powers $[f(x)]^\alpha$, $\alpha > 0$, are PSD is of the form: $f(x) = \exp\{c + \psi(x)\}$, where $\psi(x)$ is PSD and $c \in \mathbb{R}$.*

If $\psi(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) = \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$, $\sigma > 0$, and $c = -\frac{2}{\sigma}\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|$, it follows from Theorem (3.1) that K_E is PSD. This discussion suggests that, if $d_\mu(\cdot,\cdot)$ and $d_\Sigma(\cdot,\cdot)$ is (semi-)metric for $\{\boldsymbol{\mu}_i\}_{i=1}^n$ and $\{\boldsymbol{\Sigma}_i\}_{i=1}^n$ respectively, then K_μ and K_Σ can be defined as:

$$\begin{aligned} K_\mu(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j) &= \exp\left\{-\tfrac{1}{\sigma} d_\mu(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j)\right\}, \\ K_\Sigma(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j) &= \exp\left\{-\tfrac{1}{\sigma} d_\Sigma(\boldsymbol{\Sigma}_i, \boldsymbol{\Sigma}_j)\right\}, \text{ and hence} \\ K_{\mathbb{X}} &= \exp\left\{-\tfrac{1}{\sigma}[d_\mu + d_\Sigma]\right\}, \end{aligned} \tag{4}$$

where $\sigma > 0$, and $[d_\mu + d_\Sigma]$ is a (semi-)metric for the augmented space $\mathbb{X}$. In Section (4), it will be shown that, in general, d_μ is the GQD between $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$, while d_Σ is a (semi-)metric for symmetric PD covariance matrices.

3.1 Isometric embedding in a Hilbert space $\mathcal{H}$

An interesting property of the exponential function in K_E and K_G is its ability to perform an isometric embedding for $(\mathbb{R}^p, \|\cdot\|_2)$ and $(\mathbb{R}^p, \|\cdot\|_2^2)$ into a Hilbert space $\mathcal{H}$. This result is due to Theorems (1) in [17] which states that:

Theorem 2. *A necessary and sufficient condition that a separable space $\mathcal{S}$ with a semi-metric distance d, be isometrically embeddable in $\mathcal{H}$, is that the function $\exp\{-\alpha d^2\}$, $\alpha > 0$, be PSD in $\mathcal{S}$.*

Moreover, if d is a metric, then the triangle inequality is preserved through the embedding, and the new space becomes a metric space[7]. Therefore, if d_μ and d_Σ are metrics (or semi-metrics) for $\{\boldsymbol{\mu}_i\}_{i=1}^n$ and $\{\boldsymbol{\Sigma}_i\}_{i=1}^n$ respectively, then by Theorem (3.1), $K_\mu \succeq 0$ and $K_\Sigma \succeq 0$, and by Theorem (3.2), $(\{\boldsymbol{\mu}_i\}_{i=1}^n, d_\mu)$, $(\{\boldsymbol{\Sigma}_i\}_{i=1}^n, d_\Sigma)$ and $(\mathbb{X}, [d_\mu + d_\Sigma])$ are isometrically embeddable in $\mathcal{H}$.

Theorem (2) in [17], which we do not state here due to space limitations, is similar to Theorem (3.2), however it addresses the particular case of spaces with m real numbers, denoted by $\mathcal{S}_m$, and equipped with a norm function $\varphi(\mathbf{x})$, $\mathbf{x} \in \mathcal{S}_m$, and a distance function $\varphi(\mathbf{x} - \mathbf{x}')^{\frac{1}{2}}$. This theorem will be used instead of Theorem (3.2), when the Riemannian metric for symmetric PD matrices is introduced.

[7] See footnote in [17, p. 525].

4 Kernels for probability distributions

To derive d_μ and d_Σ, our discussion begins from the dual perspective for $\mathbb{X}$, or the set $\mathscr{G} = \{\mathcal{G}_i(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\}_{i=1}^n$, and the definition of K_E as an exponential function of the Euclidean distance between its input vectors. The fundamental difference here is that the elements of interests are not the vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbb{R}^p$, but rather two Gaussian distributions $\mathcal{G}_i, \mathcal{G}_j \in \mathbb{G}_p$, with $\boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_i \neq \boldsymbol{\Sigma}_j$. It follows that the Euclidean distance describing the difference between $\mathbf{x}_i$ and $\mathbf{x}_j$ needs to be replaced with a dissimilarity measure for probability distributions, and this measure should be at least a semi-metric in order to guarantee that the resulting kernel is PSD, according to Theorem (3.2).

A natural measure for the dissimilarity between probability distributions is the divergence, which by definition according to [1] and [3] is not a metric. To see this, let $\mathcal{P}$ be a family of probability distributions, and let $P_1, P_2 \in \mathcal{P}$ be defined over the same domain of events $\mathcal{E}$, then the divergence of P_2 from P_1 is:

$$\mathrm{div}(P_1, P_2) = \mathbb{E}_{p_1}\{C(\phi)\} = \int_{\mathcal{E}} p_1(x) C(\phi(x)) dx, \tag{5}$$

where $\mathrm{div}(P_1, P_2) \in [0, \infty)$, p_1, p_2 are the probability density functions of P_1 and P_2 respectively, $\phi(x) = p_1(x)/p_2(x)$ is the likelihood ratio, and C is a continuous convex function on $(0, \infty)$. Note that by definition, $\mathrm{div}(P_1, P_2) \geq 0$, and equality only holds when $P_1 = P_2$ [1]. This is equivalent to Axioms (1) & (2) of a metric, and hence $\mathrm{div}(P_1, P_2)$ is PSD. The divergence as defined in Equation (5), is not symmetric[8], since $\mathrm{div}(P_1, P_2) \neq \mathrm{div}(P_2, P_1)$. However, a possible symmetrization for the divergence can be as : $\mathrm{sdiv}(P_1, P_2) = \mathrm{div}(P_1, P_2) + \mathrm{div}(P_2, P_1)$, where sdiv preserves all the properties of a divergence as postulated by Ali–Silvey and Csiszar. Hence, sdiv is symmetric and PSD – a semi-metric – and a possible kernel for P_1 and P_2 can be defined as:

$$K_{\mathcal{P}}(P_1, P_2) = \exp\{-\tfrac{1}{\sigma}\mathrm{sdiv}(P_1, P_2)\},\ \sigma > 0. \tag{6}$$

Using Theorems (3.1) and (3.2), $K_{\mathcal{P}}$ is symmetric and PSD, and $(\mathcal{P}, \mathrm{sdiv})$ is isometrically embeddable in $\mathcal{H}$. Note that $K_{\mathcal{P}}$ is in the same spirit of the exponential kernel K_E as explained above. In addition, $K_{\mathcal{P}}$ is valid for any symmetric divergence measure from the class of Ali–Silvey or f–divergence [3], and hence it is valid for any probability distribution. It is also important to note that the kernel $K_{\mathcal{P}}$ is not the only kernel for probability distributions, and other kernels were proposed in the work of [6, 4, 11].

4.1 The case of Gaussian densities

We now consider the particular case of Gaussian densities under some classical symmetric divergence measures such as the symmetric KL divergence, or Jeffreys

[8] Depending on the choice of $C(\cdot)$ in (5) and its parametrization, one can derive symmetric divergence measures, see [1] for examples.

divergence d_J, the Bhattacharyya divergence d_B, and the Hellinger distance d_H. For $\mathcal{G}_1, \mathcal{G}_2 \in \mathbb{G}_p$, Jeffreys divergence d_J can be expressed as:

$$d_J(\mathcal{G}_1, \mathcal{G}_2) = \tfrac{1}{2}\mathbf{u}^\top \boldsymbol{\Psi}\mathbf{u} + \tfrac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\} - p, \tag{7}$$

where $\boldsymbol{\Psi} = (\boldsymbol{\Sigma}_1^{-1} + \boldsymbol{\Sigma}_2^{-1})$, and $\mathbf{u} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. The Bhattacharyya divergence d_B and the Hellinger distance d_H are both derived from the Bhattacharyya coefficient ρ, which is a measure of similarity between probability distributions:

$$\rho(\mathcal{G}_1, \mathcal{G}_2) = |\boldsymbol{\Gamma}|^{-\frac{1}{2}}|\boldsymbol{\Sigma}_1|^{\frac{1}{4}}|\boldsymbol{\Sigma}_2|^{\frac{1}{4}}\exp\{-\tfrac{1}{8}\mathbf{u}^\top\boldsymbol{\Gamma}^{-1}\mathbf{u}\},$$

where $\boldsymbol{\Gamma} = (\frac{1}{2}\boldsymbol{\Sigma}_1 + \frac{1}{2}\boldsymbol{\Sigma}_2)$. The Hellinger distance can be obtained from ρ as $d_H(\mathcal{G}_1, \mathcal{G}_2) = \sqrt{2[1 - \rho(\mathcal{G}_1, \mathcal{G}_2)]}$, while $d_B(\mathcal{G}_1, \mathcal{G}_2) = \log[\rho(\mathcal{G}_1, \mathcal{G}_2)]$ is defined as:

$$d_B(\mathcal{G}_1, \mathcal{G}_2) = \tfrac{1}{8}\mathbf{u}^\top\boldsymbol{\Gamma}^{-1}\mathbf{u} + \tfrac{1}{2}\ln\left\{\frac{|\boldsymbol{\Gamma}|}{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}|\boldsymbol{\Sigma}_2|^{\frac{1}{2}}}\right\}. \tag{8}$$

Kullback [9] notes that d_J is positive and symmetric but violates the triangle inequality. Similarly, Kailath [7] notes that d_B is positive and symmetric but violates the triangle inequality, while d_H meets all metric Axioms. Using the kernel definition in (6), it is straight forward to define the following kernels:

$$K_J(\mathcal{G}_1, \mathcal{G}_2) = \exp\{-\tfrac{1}{\sigma}d_J(\mathcal{G}_1, \mathcal{G}_2)\},\ \sigma > 0, \tag{9}$$

$$K_H(\mathcal{G}_1, \mathcal{G}_2) = \exp\{-\tfrac{1}{\sigma}d_H(\mathcal{G}_1, \mathcal{G}_2)\},\ \sigma > 0,\ \text{and} \tag{10}$$

$$K_B(\mathcal{G}_1, \mathcal{G}_2) = \exp\{d_B(\mathcal{G}_1, \mathcal{G}_2)\} = \rho(\mathcal{G}_1, \mathcal{G}_2). \tag{11}$$

We note that [8] have proposed the Bhatacharyya kernel $\rho(\mathcal{G}_1, \mathcal{G}_2)$ and confirm that it is PSD through the product probability kernel (PPK). In contrary, [12] have proposed the KL kernel $K_J(\mathcal{G}_1, \mathcal{G}_2)$ and claim, without justification, that it is not PSD. Since d_J and d_B are semi-metrics, and d_H is a metric, then using Theorems (3.1) and (3.2), K_J, K_H and K_B are symmetric and PSD kernels, and $(\mathbb{X}, d_J)$, $(\mathbb{X}, d_B)$, and $(\mathbb{X}, d_H)$ are isometrically embeddable in $\mathcal{H}$.

4.2 A close look at d_J and d_B

Kullback [9, pp. 6,7] describes $d_J(\mathcal{G}_1, \mathcal{G}_2)$ in Equation (7) as a sum of two components, one due to the difference in means weighted by the covariance matrices (the first term), and the other due to the difference in variances and covariances (the second term). Note that this explanation is also valid for $d_B(\mathcal{G}_1, \mathcal{G}_2)$ in Equation (8). Recalling $K_{\mathbb{X}}$ from Equation (4), then d_μ and d_Σ can be characterized as follows. The first term in Equations (7) and (8) is equivalent to the GQD, up to a constant and a square root – hence both terms are semi-metrics. If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, then:

$$\left.\begin{array}{l} d_J(\mathcal{G}_1, \mathcal{G}_2) = \mathbf{u}^\top\boldsymbol{\Psi}\mathbf{u}, \\ d_B(\mathcal{G}_1, \mathcal{G}_2) = \mathbf{u}^\top\boldsymbol{\Gamma}^{-1}\mathbf{u}. \end{array}\right\} d_\mu \tag{12}$$

The second term in Equations (7) and (8) is a discrepancy measure between two covariance matrices that is independent from $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$. If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$ then:

$$\left.\begin{array}{l} d_J(\mathcal{G}_1, \mathcal{G}_2) = \mathrm{tr}\{\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2 + \boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1\} - p, \\ d_B(\mathcal{G}_1, \mathcal{G}_2) = \ln\left\{ |\boldsymbol{\Gamma}||\boldsymbol{\Sigma}_1|^{-\frac{1}{2}}|\boldsymbol{\Sigma}_2|^{-\frac{1}{2}} \right\}, \end{array}\right\} d_\Sigma \tag{13}$$

which define two dissimilarity measures between $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$, and both measures are semi-metrics.

4.3 A metric for symmetric PD matrices

The factorisable nature of $K_{\mathbb{X}}$, and the decomposition of $d_J(\mathcal{G}_1, \mathcal{G}_2)$ and $d_B(\mathcal{G}_1, \mathcal{G}_2)$ into two difference components, where the second term is independent from $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, allows us to introduce a metric for symmetric PD matrices that can be used instead of the semi-metrics in Equation (13).

A symmetric PD matrix is a geometric object, and the space of all symmetric PD matrices, denoted by $\mathbb{S}_{++}^{p\times p}$, is a differentiable manifold in which each point $\mathbf{A} \in \mathbb{S}_{++}^{p\times p}$ has a tangent space $\mathcal{T}_{\mathbf{A}}(\mathbb{S}_{++}^{p\times p})$ that is endowed with an inner product, or a Riemmanian metric $\langle\cdot,\cdot\rangle_{\mathbf{A}}$, on the elements of the tangent space. The dimensionality of $\mathbb{S}_{++}^{p\times p}$ and its tangent space is $p(p+1)/2$. Due to the inner product $\langle\cdot,\cdot\rangle_{\mathbf{A}}$, the tangent space for $\mathbb{S}_{++}^{p\times p}$ is a finite dimensional Euclidean space.

The Riemannian metric, by default, respects the geometry of $\mathbb{S}_{++}^{p\times p}$, which is unlike the semi-metrics in (13) that are just derived from the divergence measures $d_J(\mathcal{G}_1, \mathcal{G}_2)$ and $d_B(\mathcal{G}_1, \mathcal{G}_2)$, and unaware of the geometry of $\mathbb{S}_{++}^{p\times p}$. If $d_{\mathcal{R}}$ is the Riemannian metric for $\mathbb{S}_{++}^{p\times p}$, then d_Σ in Equation (4) can be replaced with $d_{\mathcal{R}}$, and hence $K_{\mathbb{X}}$ can be redefined as follows:

$$\begin{aligned} K_{\mathbb{X}} &= K_\mu(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2) K_{\mathcal{R}}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2), && (14) \\ &= \exp\{-\tfrac{1}{\sigma} d_\mu\} \exp\{-\tfrac{1}{\sigma} d_{\mathcal{R}}\}, \\ &= \exp\{-\tfrac{1}{\sigma}[d_\mu + d_{\mathcal{R}}]\},\ \sigma > 0. && (15) \end{aligned}$$

where $d_{\mathcal{R}}$ is the distance between the two matrices $\{\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathbb{S}_{++}^{d\times d}\}$ defined as :

$$d_{\mathcal{R}}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \mathrm{tr}\{\ln^2 \boldsymbol{\Lambda}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2)\}^{\frac{1}{2}}, \tag{16}$$

and $\boldsymbol{\Lambda}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is the solution of a generalized eigenvalue problem (GEP): $\boldsymbol{\Sigma}_1\mathbf{V} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_2\mathbf{V}$. The metric $d_{\mathcal{R}}$ was first derived by C. Rao [16], and latter analyzed by Atkinson and Mitchel [2] [9], while independently derived by Förstner and Moonen in [5]. Note that $d_{\mathcal{R}}$ is invariant to inversion and to affine transformations of the coordinate system. Since $d_{\mathcal{R}}$ is induced by a norm on $\mathcal{T}(\mathbb{S}_{++}^{p\times p})$, then using Theorem (3.1) and Theorem(2) in [17], $K_{\mathcal{R}}$ is PSD, and $(\mathcal{T}_{\mathbf{A}}(\mathbb{S}_{++}^{p\times p}), d_{\mathcal{R}})$ is isometrically embeddable in $\mathcal{H}$, for all $\mathbf{A} \in \mathbb{S}_{++}^{p\times p}$.

[9] See their affiliated references

5 Relaxed kernels for the augmented space $\mathbb{X}$

Besides the Jeffreys kernel K_J, the Hellinger kernel K_H, and the Bhattacharyya kernel K_B in Equations (9), (10) and (11) respectively, we define two new kernels for the space $\mathbb{X}$ based on the metric $d_{\mathcal{R}}$:

$$K_{J\mathcal{R}}(\mathcal{G}_1, \mathcal{G}_2) = \exp\{-\tfrac{1}{\sigma} d_{J\mathcal{R}}(\mathcal{G}_1, \mathcal{G}_2)\}, \quad \text{and} \tag{17}$$

$$K_{B\mathcal{R}}(\mathcal{G}_1, \mathcal{G}_2) = \exp\{-\tfrac{1}{\sigma} d_{B\mathcal{R}}(\mathcal{G}_1, \mathcal{G}_2)\}, \quad \text{where} \tag{18}$$

$$d_{J\mathcal{R}}(\mathcal{G}_1, \mathcal{G}_2) = (\mathbf{u}^\top \boldsymbol{\Psi} \mathbf{u})^{\frac{1}{2}} + d_{\mathcal{R}}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2),$$

$$d_{B\mathcal{R}}(\mathcal{G}_1, \mathcal{G}_2) = (\mathbf{u}^\top \boldsymbol{\Gamma}^{-1} \mathbf{u})^{\frac{1}{2}} + d_{\mathcal{R}}(\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2),$$

$$\boldsymbol{\Psi} \succ 0, \quad \boldsymbol{\Gamma}^{-1} \succ 0, \quad \text{and} \quad \sigma > 0.$$

The positive definiteness of $\boldsymbol{\Psi}$ and $\boldsymbol{\Gamma}^{-1}$, and the square root on the quadratic terms of $d_{J\mathcal{R}}$ and $d_{B\mathcal{R}}$, assure that the quadratic terms are metrics. If $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$, then $d_{J\mathcal{R}}$ and $d_{B\mathcal{R}}$ will yield the Riemannian metric $d_{\mathcal{R}}$, and hence, $K_{J\mathcal{R}}$ and $K_{B\mathcal{R}}$ will be equal to $K_{\mathcal{R}}$. If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$, then $d_{J\mathcal{R}}$ and $d_{B\mathcal{R}}$ will yield the GQD. If $\boldsymbol{\Sigma} = \mathbf{I}$, the GQD will be equal to the Euclidean distance, and $K_{J\mathcal{R}}$ and $K_{B\mathcal{R}}$ will yield the original exponential kernel K_E.

Similar to K_E and K_G, the relaxed kernels K_J, K_H, K_B, $K_{J\mathcal{R}}$ and $K_{B\mathcal{R}}$ rely on the distance between the 2-tuples $(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$. Moreover, they all provide an isometric embedding for the space $\mathbb{X}$, and the difference between these embeddings is due the metric or semi-metric defining each kernel. While d_J and d_B are semi-metrics, d_H, $d_{J\mathcal{R}}$ and $d_{B\mathcal{R}}$ are metrics. Since Axioms (3) & (5) do not hold for semi-metrics, it follows that d_J and d_B will not preserve the relative geometry between the elements in $\mathbb{R}^p$, and that between the elements in $\mathbb{S}_{++}^{p\times p}$. Although d_H is a metric, it relies on a semi-metric for covariances matrices, which is not the case for $d_{J\mathcal{R}}$ and $d_{B\mathcal{R}}$.

6 Related work

Our research work parallels a stream of ideas that consider distances (or similarities) between two subspaces, tangent spaces, or sets of vectors, instead of the direct distance (or similarity) between points. In the context of learning over sets of vectors (SOV's), [20] propose a general learning approach within the kernel framework. For two SOV's, their kernel is based on the principal angles between two subspaces, each spanned by one of the two SOV's. In [8], each SOV is a bag of pixels representing one image. Each SOV is modelled as a Gaussian distribution, and the Bhattacharyya kernel K_B is used with SVMs to classify the images. Similarly, in [12] each SOV is a bag of features representing one multimedia object (an image or an audio signal), and modelled as Gaussian distribution. However, instead of K_B, they use the KL kernel K_J with SVMs to classify the multimedia objects.

Table 1. Specifications of the data sets used in the experiments. The number of classes, samples, and attributes are denoted by c, n, and p respectively.

Data set	c	n	p	Data set	c	n	p
Balance	3	625	4	NewThyroid	3	215	5
Bupa	2	345	6	Pima	2	768	8
Glass	6	214	9	Segment	7	2310	18
Iris	3	150	4	Sonar	2	208	60
Lymphography	4	148	18	WDBC	2	569	30
Monks–1	2	556	6	Wine	3	178	13
Monks–2	2	601	6	Yeast	10	1484	6
Monks–3	2	554	6				

7 Experimental results

We validate the proposed relaxed kernels in the context of unsupervised learning using spectral clustering (SC) algorithms. Here we compare the performance of 1) the standard k-means algorithm, 2) SC according to the version of [14]–as described in [10]–using the exponential kernel K_E, and 3) SC over the augmented space $\mathbb{X}$ using four (4) different kernels: the KL kernel K_J [12], the Bhattacharyya kernel K_B [8], the Hellinger kernel K_H, and the proposed kernel $K_{B\mathcal{R}}$. Although our experiments included $K_{J\mathcal{R}}$ as well, we found that the results of $K_{J\mathcal{R}}$ and $K_{B\mathcal{R}}$ are very close to each other, and hence we show only the results for $K_{B\mathcal{R}}$. This shows that the main difference between d_J and d_B are the semi-metrics for covariance matrices in Equation (13). The parameter σ for K_E, K_J, K_B, K_H and $K_{B\mathcal{R}}$ was selected using a simple quantile based approach[10]. In all our experiments, the regularization parameter $\gamma = 1$. Although we do not focus on selecting the *best* γ values, it nevertheless shows that, under this uniform γ assumption, the local Gaussian assumption typically shows significantly better results.

All algorithms were run on 15 data sets from the UCI machine learning repository [13], shown in Table (1). Clustering accuracy was measured using the Hungarian score of [21] [11]. The performance of each algorithm was averaged over 30 runs with different initializations. Since the number of classes of the UCI data sets is given, we assumed that the number of clusters is known. Before proceeding to the results, it is important to emphasize that selecting the best parameter values for k, σ, γ and the number of clusters, is largely a model selection issue, and hence, it should not be confounded with verification of the effectiveness of the local Gaussian modelling premise.

Columns two and three in Table (2) show the results for k-Means and SC with K_E on the original data set $\mathcal{X}$. Columns four to seven in Table (2) show the results of SC over the augmented data set $D_A = \{(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)\}_{i=1}^n$ with the 4 different relaxed kernels. Due to space limitaitons, we do not show the results

[10] The approach was suggested in Alex Smola's blog: `http://blog.smola.org/page/2`
[11] See [21] for more details.

Table 2. Clustering accuracy for k-Means, SC with K_E, and SC over $D_A = \{(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)\}_{i=1}^n$ with K_J, K_B, K_H, and $K_{B\mathcal{R}}$.

Data set	k-Means	K_E	K_J	K_B	K_H	$K_{B\mathcal{R}}$
Balance	51.1 (3.2)	53.6 (4.5)	60.0 (3.4)	61.2 (3.5)	60.3 (0.2)	**64.9 (4.9)**
Bupa	55.1 (0.1)	57.1 (0.2)	62.9 (0.1)	62.3 (0.18)	61.1 (0.07)	**64.3 (0.18)**
Glass	51.3 (3.4)	50.8 (2.0)	52.5 (2.8)	52.4 (3.3)	52.2 (2.9)	**53.8 (4.6)**
Iris	84.5 (12.0)	93.5 (7.1)	93.1 (18.1)	**94.4 (9.2)**	93.4 (5.4)	93.3 (10.4)
Lymphography	48.0 (6.1)	52.5 (5.7)	65.9 (2.1)	**69.6 (2.9)**	67.6 (5.3)	65.0 (6.0)
Monks–1	64.6 (5.4)	66.4 (0.1)	62.0 (2.1)	62.7 (1.0)	67.6 (0.02)	**69.4 (0.04)**
Monks–2	51.6 (2.0)	53.1 (3.0)	**65.3 (0.1)**	**65.3 (0.02)**	**65.4 (1.7)**	**65.7 (0.1)**
Monks–3	63.4 (4.2)	65.7 (0.1)	**79.9 (0.02)**	**79.9 (0.02)**	**79.9 (0.02)**	**79.9 (0.02)**
NewThyroid	78.0 (9.7)	75.8 (0.1)	79.9 (1.5)	80.3 (0.4)	86.5 (3.2)	**91.1 (2.5)**
Pima	66.0 (0.1)	64.7 (0.2)	**68.2 (0.1)**	67.8 (0.16)	67.9 (0.02)	67.5 (0.05)
Segment	51.5 (8.1)	65.4 (5.1)	62.7 (13.7)	62.9 (6.5)	67.7 (3.6)	**69.1 (5.7)**
Sonar	54.5 (0.7)	54.8 (4.2)	63.2 (3.1)	**64.2 (2.4)**	63.1 (2.1)	**64.3 (2.4)**
WDBC	85.4 (0.1)	84.0 (0.1)	92.7 (5.4)	93.6 (0.08)	94.2 (0.1)	**95.6 (0.1)**
Wine	67.8 (5.1)	67.8 (7.0)	**90.4 (4.9)**	88.1 (8.9)	88.7 (0.3)	88.2 (6.3)
Yeast	34.2 (1.3)	42.7 (2.8)	**46.9 (1.8)**	45.5 (1.8)	45.0 (2.2)	45.4 (2.5)

for SC over the augmented data set $D_A = \{(\mathbf{x}_i, \mathbf{A}_i)\}_{i=1}^n$. It can be seen that for most of the cases, the performance of SC over the augmented data sets outperforms the standard SC and the k-Means algorithms. More specifically, the performance of SC over $D_A = \{(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)\}_{i=1}^n$, is consistently better than k-Means and the standard SC, which is due to the smoothing included in defining the 2-tuple $(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)$. In terms of kernels over $D_A = \{(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\Sigma}}_i)\}_{i=1}^n$, K_H and $K_{B\mathcal{R}}$ are usually better than K_J and K_B, and at least, very close to their performance. This emphasizes the role of the (semi-)metric defining each kernel.

8 Conclusion

We relax the global Gaussian assumption of the Euclidean distance in the exponential kernel K_E. The relaxation anchors a Gaussian distribution on the local neighbourhood of each point in the data set, resulting in the augmented data space $\mathbb{X}$. Based on convolution kernels, divergence measures of probability distributions, and Riemannian metrics for symmetric PD matrices, we propose a set of kernels for the space $\mathbb{X}$, and show using preliminary experiments that the local Gaussian assumption significantly outperforms the global one. Since all our approach described here is unsupervised, a main future research direction is to investigate the usefulness of this approach in supervised and semi-supervised learning tasks.

References

1. Ali, S.M., Silvey, S.D.: A general class of coefficients of divergence of one distribution from another. J. of the Royal Statistical Society. *Series B* 28(1), 131–142

(1966)
2. Atkinson, C., Mitchell, A.F.S.: Rao's distance measure. The Indian J. of Statistics, Series A 43(3), 345–365 (1945)
3. Csiszár, I.: Information–type measures of difference of probability distributions and indirect observations. Studia Scientiarium Mathematicarum Hungarica 2, 299–318 (1967)
4. Cuturi, M., Fukumizu, K., Vert, J.P.: Semigroup kernels on measures. JMLR 6, 1169–1198 (2005)
5. Förstner, W., Moonen, B.: A metric for covariance matrices. Tech. rep., Dept. of Geodesy and Geo–Informatics, Stuttgart University (1999)
6. Hein, M., bousquet, O.: Hilbertian metrics and positive definite kernels on probability measures. In: Proc. of AISTATS. pp. 136–143 (2005)
7. Kailath, T.: The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. on Communication Technology 15(1), 52–60 (1967)
8. Kondor, R., Jebara, T.: A kernel between sets of vectors. In: ACM Proc. of ICML (2003)
9. Kullback, S.: Information Theory and Statistics – Dover Edition. Dover, New York (1997)
10. Luxburg, U.v.: A tutotrial on spectral clustering. Tech. Rep. TR–149, Max Plank Institute for Biological Cybernetics (2006)
11. Martins, A., Smith, N., Xing, E., Aguiar, P., Figueiredo, M.: Nonextensive information theoretic kernels on measures. JMLR 10, 935–975 (2009)
12. Moreno, P., Ho, P., Vasconcelos, N.: A Kullback–Leibler divergence based kernel for svm classification in multimedia applications. In: NIPS 16 (2003)
13. Newman, D., Hettich, S., Blake, C., Merz, C.: UCI Repository of Machine Learning Databases (1998), `www.ics.uci.edu/~mlearn/MLRepository.html`
14. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: NIPS 14. pp. 849–856. MIT Press (2002)
15. Pennec, X., Fillard, P., Ayache, N.: A Riemannian framework for tensor computing. Int. J. Computer Vision 66(1), 41–66 (2006)
16. Rao, C.R.: Information and the accuracy attainable in the estimation of statistical parameters. Bull. Calcutta Math. Soc. (58), 326–337 (1945)
17. Schoenberg, I.: Metric spaces and positive definite functions. Trans. of the American Mathematical Society 44(3), 522–536 (1938)
18. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press (2004)
19. Vincent, P., Bengio, Y.: Manifold Parzen windows. In: NIPS 15. pp. 825–832. MIT Press (2003)
20. Wolf, L., Shashua, A.: Learning over sets using kernel principal angles. JMLR 4, 913–931 (Dec 2003)
21. Zha, H., Ding, C., Gu, M., He, X., Simon, H.: Spectral relaxation for k–means clustering. In: NIPS 13. MIT Press (2001)