

# Normalized Kernels as Similarity Indices

Julien Ah-Pine

▶ **To cite this version:**

Julien Ah-Pine. Normalized Kernels as Similarity Indices. 14th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2010), Jun 2010, Hyderabad, India. pp.362 - 373, 10.1007/978-3-642-13672-6_36 . hal-01504523

**HAL Id: hal-01504523**
**https://hal.science/hal-01504523**

Submitted on 10 Apr 2017

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Normalized kernels as similarity indices

Julien Ah-Pine

Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
julien.ah-pine@xrce.xerox.com

**Abstract.** Measuring similarity between objects is a fundamental issue for numerous applications in data-mining and machine learning domains. In this paper, we are interested in kernels. We particularly focus on kernel normalization methods that aim at designing proximity measures that better fit the definition and the intuition of a similarity index. To this end, we introduce a new family of normalization techniques which extends the cosine normalization. Our approach aims at refining the cosine measure between vectors in the feature space by considering another geometrical based score which is the mapped vectors' norm ratio. We show that the designed normalized kernels satisfy the basic axioms of a similarity index unlike most unnormalized kernels. Furthermore, we prove that the proposed normalized kernels are also kernels. Finally, we assess these different similarity measures in the context of clustering tasks by using a kernel PCA based clustering approach. Our experiments employing several real-world datasets show the potential benefits of normalized kernels over the cosine normalization and the Gaussian RBF kernel.

**Key words:** Kernels normalization, similarity indices, kernel PCA based clustering.

## 1 Introduction

Measuring similarity between objects is a fundamental issue for numerous applications in data-mining and machine learning domains such as in clustering or in classification tasks. In that context, numerous recent approaches that tackle the latter tasks are based on kernels (see for example [1]). Kernels are special dot products considered as similarity measures. They are popular because they implicitly map objects initially represented in an input space, to a higher dimensional space, called the feature space. The so-called kernel trick relies on the fact that they represent dot products of mapped vectors without having to explicitly represent the latter in the feature space. From a practical standpoint, kernel methods allow one to deal with data that are not easy to linearly separate in the input space. In such cases, any clustering or classification method that makes use of dot products in the input space is limited. By mapping the data to a higher dimensional space, those methods can thus perform much better. Consequently, many other kinds of complex objects can be efficiently treated by

using kernel methods. We have mentioned previously that kernels are generally introduced as similarity measures but as underlined in [2], dot products in general do not necessarily fit one's intuition of a similarity index. Indeed, one could find in the literature several axioms that clarify the definition of a similarity index and in that context, any kernel does not necessarily satisfy all of them. As an example, one of these conditions that a dot product, and thus a kernel, does not always respect, is the maximal self-similarity axiom which states that the object to which any object should be the most similar, is itself.

In this paper, we are interested in designing kernels which respect the basic axioms of a geometrical based similarity index. In that context, kernels normalization methods are useful. Basically, the most common way to normalize a kernel so as to have a similarity index, is to apply the cosine normalization. In that manner, maximal self-similarity for instance, is respected unlike for unnormalized kernels. In this work we propose a new family of kernel normalization methods that generalizes the cosine normalization. Typically, the cosine normalization leads to similarity measures between vectors that are based upon their angular measure. Our proposal goes beyond the cosine measure by refining the latter score by using another geometrical based measure which relies on the vectors' norm ratio in the feature space. We give the following example in order to motivate such normalized kernels. Let us take two vectors which are positively colinear in the feature space. In that case, their cosine measure is 1. However, if their norms are not the same ones therefore, we cannot conclude that these two mapped vectors are identical. Accordingly, their similarity measure should be lower than 1. Unlike the cosine normalization, the normalization approaches that we introduce in this paper aim at taking into consideration this point.

The rest of this paper is organized as follows. In section 2, we formally introduce new normalization methods for kernels. Then, in section 3, we give several properties of the resulting normalized kernels in the context of similarity indices. We show that using normalization methods allows one to make any kernel satisfy the basic axioms of a similarity index. Particularly, we prove that these kernel normalizations define metrics. In other words, we show that normalized kernels are kernels. In section 4, we illustrate the benefits of our proposal in the context of clustering tasks. The method we use in that regard, relies on kernel PCA based $k$-means clustering which can be understood as a combination between kernel PCA [3] and $k$-means via PCA [4]. This two step approach is a spectral clustering like algorithm. Using several datasets from the UCI ML repository [5], we show that different normalizations can better capture the proximity relationships of objects and improve the clustering results of the cosine measure and of another widely used normalized kernel, the Gaussian Radial Basis Function (RBF) kernel. We finally conclude and sketch some future works in section 5.

## 2 Kernels and normalization methods

We first recall some basic definitions about kernel functions and their cosine normalization. We then introduce our new kernel normalization methods.

### 2.1 Kernel definition and the cosine normalization

Let denote $\mathcal{X}$ the set of objects, represented in an input space, that we want to analyze.

**Definition 1 ((Positive semi-definite) Kernel).** *A function $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a positive semi-definite kernel if it is symmetric, that is, $K(\mathbf{x}, \mathbf{y}) = K(\mathbf{y}, \mathbf{x})$ for any two objects $\mathbf{x}, \mathbf{y}$ in $\mathcal{X}$ and positive semi-definite, that is:*

$$\sum_{i=1}^{n} \sum_{i'=1}^{n} c_i c_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'}) \geq 0 \tag{1}$$

*for any $n > 0$, any choice of $n$ objects $\mathbf{x}_1, \ldots, \mathbf{x}_n$ in $\mathcal{X}$ and any choice of any numbers $c_1, \ldots, c_n$ in $\mathbb{R}$.*

In the sequel, we will simply use the term kernel instead of positive semi-definite kernel. We have the following well-known property.

**Theorem 1.** *For any kernel $K$ on an input space $\mathcal{X}$, there exists a Hilbert space $\mathcal{F}$, called the feature space, and a mapping $\phi : \mathcal{X} \rightarrow \mathcal{F}$ such that for any two objects $\mathbf{x}, \mathbf{y}$ in $\mathcal{X}$:*

$$K(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle \tag{2}$$

*where $\langle ., . \rangle$ is the Euclidean dot product.*

When the objects of $\mathcal{X}$ are vectors represented in an input space which is an Euclidean space then, we can mention the following two well-known types of kernel functions:

- Polynomial kernels: $K_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + c)^d$, with $d \in \mathbb{N}$ and $c \geq 0$.
- Gaussian RBF kernels: $K_g(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$, where $\|.\|$ is the Euclidean norm and $\sigma > 0$.

After having recalled basics about kernels, we recall the definition of the cosine normalization of a kernel $K$, denoted $K^0$.

$$K^0(\mathbf{x}, \mathbf{y}) = \frac{K(\mathbf{x}, \mathbf{y})}{\sqrt{(K(\mathbf{x}, \mathbf{x}) K(\mathbf{x}, \mathbf{y}))}} \tag{3}$$

Since we have $K(\mathbf{x}, \mathbf{x}) = \|\phi(\mathbf{x})\|^2$, it is easy to see that:

$$K^0(\mathbf{x}, \mathbf{y}) = \langle \frac{\phi(\mathbf{x})}{\|\phi(\mathbf{x})\|}, \frac{\phi(\mathbf{y})}{\|\phi(\mathbf{y})\|} \rangle$$

Moreover, we have $K^0(\mathbf{x}, \mathbf{x}) = 1$ for all objects in $\mathcal{X}$. This means that the objects in the feature space are projected on an unit hypersphere. In addition, we have the following geometrical interpretation from the feature space representation viewpoint:

$$K^0(\mathbf{x}, \mathbf{y}) = \cos(\theta(\phi(\mathbf{x}), \phi(\mathbf{y}))) \tag{4}$$

with $\theta(\phi(\mathbf{x}), \phi(\mathbf{y}))$ being the angular measure between the vectors in the feature space. In order to simplify the notations, we will denote $\theta$ for $\theta(\mathbf{x}, \mathbf{y})$ and $\cos\theta$ for $\cos(\theta(\phi(\mathbf{x}), \phi(\mathbf{y})))$, thereafter.

### 2.2 Kernel normalization of order $t$

The main purpose of this paper is to introduce a new family of kernel normalization approaches that generalizes the cosine normalization. Our approach amounts to integrating another geometrical based measure that allows us to refine the cosine measure $K^0$. This additional feature is related to the difference between the norms of the two vectors in the feature space.

These normalization procedures involve generalized mean (also known as power mean) operators which generalize the classical arithmetic mean. Given a sequence of $p$ values $\{a_i\}_{i=1}^{p} = \{a_1, a_2, \ldots, a_p\}$, the generalized mean with exponent $t$ is given by:

$$\mathcal{M}^t(a_1, \ldots, a_p) = \left[\frac{1}{p}\sum_{i=1}^{p} a_i^t\right]^{\frac{1}{t}} \tag{5}$$

Famous particular cases of (5) are given by $t = -1$, $t \to 0$ and $t = 1$ which are respectively the harmonic, geometric and arithmetic means.

**Definition 2 (Kernels normalization of order $t > 0$).** *Given a kernel function $K$, the normalized kernel of order $t > 0$ for any two objects $\mathbf{x}$ and $\mathbf{y}$ of $\mathcal{X}$, is denoted $K^t(\mathbf{x}, \mathbf{y})$ and is defined as follows:*

$$K^t(\mathbf{x}, \mathbf{y}) = \frac{K(\mathbf{x}, \mathbf{y})}{\mathcal{M}^t(K(\mathbf{x}, \mathbf{x}), K(\mathbf{y}, \mathbf{y}))} \tag{6}$$

Similarly to the cosine normalization, $K^t(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}$. As a result, those normalization methods also amount to projecting the objects from the feature space to an unit hypersphere. However, this family of normalized kernels goes beyond the cosine measure since it extends the latter which actually corresponds to the limit case $t \to 0$.

In order to better interpret such measures, let us equivalently formulate $K^t(\mathbf{x}, \mathbf{y})$ with respect to the following norm ratio measures, $\frac{\|\phi(\mathbf{x})\|}{\|\phi(\mathbf{y})\|}$ and $\frac{\|\phi(\mathbf{y})\|}{\|\phi(\mathbf{x})\|}$. We can easily show that:

$$K^t(\mathbf{x}, \mathbf{y}) = \frac{\cos\theta}{\mathcal{M}^t\left(\frac{\|\phi(\mathbf{x})\|}{\|\phi(\mathbf{y})\|}, \frac{\|\phi(\mathbf{y})\|}{\|\phi(\mathbf{x})\|}\right)} \tag{7}$$

This formulation expresses $K^t(\mathbf{x}, \mathbf{y})$ according to geometrical based measures. However, let us introduce the following notation as well:

$$\gamma(\phi(\mathbf{x}), \phi(\mathbf{y})) = \max\left(\frac{\|\phi(\mathbf{x})\|}{\|\phi(\mathbf{y})\|}, \frac{\|\phi(\mathbf{y})\|}{\|\phi(\mathbf{x})\|}\right) = \frac{\max(\|\phi(\mathbf{x})\|, \|\phi(\mathbf{y})\|)}{\min(\|\phi(\mathbf{x})\|, \|\phi(\mathbf{y})\|)} \tag{8}$$

$\gamma(\phi(\mathbf{x}), \phi(\mathbf{y}))$ lies within $[1, +\infty[$ and is related to the difference between the norm measures of $\phi(\mathbf{x})$ and $\phi(\mathbf{y})$. $\gamma(\phi(\mathbf{x}), \phi(\mathbf{y})) = 1$ means that $\|\phi(\mathbf{x})\| = \|\phi(\mathbf{y})\|$ and the greater the difference between the norms' value, the higher $\gamma(\phi(\mathbf{x}), \phi(\mathbf{y}))$.

Similarly to the angular measure, we will denote $\gamma$ for $\gamma(\phi(\mathbf{x}), \phi(\mathbf{y}))$ in order to simplify the notations. Using $\gamma$, we have the different formulations below:

$$K^t(\mathbf{x}, \mathbf{y}) = K^t(\theta, \gamma) = \frac{\cos\theta}{\mathcal{M}^t(\gamma, \gamma^{-1})} = \cos\theta \left( \frac{2^{1/t}\gamma}{(1+\gamma^{2t})^{1/t}} \right) \tag{9}$$

The latter relation expresses $K^t(\mathbf{x}, \mathbf{y})$ as a multiplication between two factors. On the one hand, we have the cosine index $\cos\theta$ and on the other hand, we have the following term which is only dependent on $\gamma$, $\left( \frac{2^{1/t}\gamma}{(1+\gamma^{2t})^{1/t}} \right)$.

Following (9), we observe that $\forall\theta : \cos\theta \in [-1, 1]$ and $\forall t > 0, \forall\gamma \geq 1 : \left( \frac{2^{1/t}\gamma}{(1+\gamma^{2t})^{1/t}} \right) \in ]0, 1]$. As a result, one can see that $\forall t > 0, \forall(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^2 : K^t(\mathbf{x}, \mathbf{y}) \in [-1, 1]$.

In what follows, we detail the roles the geometrical parameters $\theta$ and $\gamma$ play with respect to the introduced normalized kernels of order $t$. First, since $\forall(\phi(\mathbf{x}), \phi(\mathbf{y})) \in \mathcal{F}^2 : \gamma \geq 1$; $K^t(\theta, \gamma)$ is clearly a monotonically increasing function with respect to $\cos\theta$[1]. Then, using (9), we can better underline the effect of the norm ratio $\gamma$ on $K^t(\theta, \gamma)$. Indeed, by computing the first derivative with respect to this parameter, we obtain:

$$\frac{\partial K^t}{\partial \gamma} = \cos\theta \left[ \frac{2^{1/t}(1+\gamma^{2t})^{1/t} \left(1 - \frac{2\gamma^{2t}}{(1+\gamma^{2t})}\right)}{(1+\gamma^{2t})^{2/t}} \right] \tag{10}$$

With the following conditions, $\gamma \geq 1$, $t > 0$, one can verify that in the numerator of the second term of (10), the first factor is positive but the second one, $\left(1 - \frac{2\gamma^{2t}}{(1+\gamma^{2t})}\right)$, is negative since $\frac{2\gamma^{2t}}{(1+\gamma^{2t})} \in [1, 2[$. Consequently, the sign of $\frac{\partial K^t}{\partial \gamma}$ is the same as $-\cos\theta$. Therefore, for $t > 0$, $K^t(\theta, \gamma)$ is monotonically decreasing with respect to $\gamma$ providing that $\cos\theta > 0$ whereas, $K^t(\theta, \gamma)$ is monotonically increasing with respect to $\gamma$ as long as $\cos\theta < 0$ (see Fig. 1).

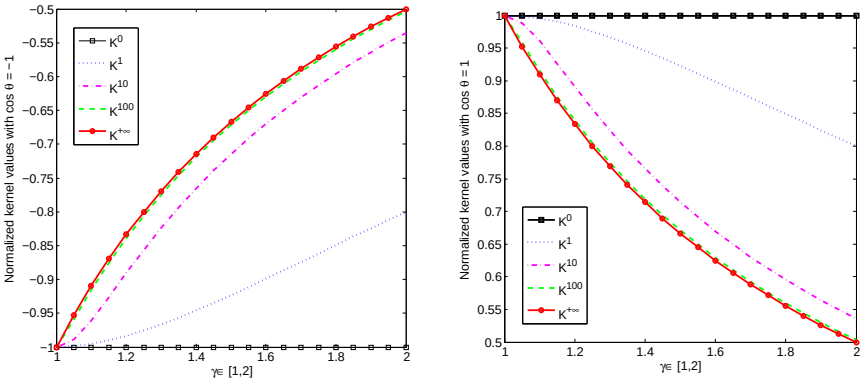
Intuitively, when $t > 0$, the norm ratio measure aims at refining the cosine measure considering that the latter is less and less "reliable" for measuring proximity, as the difference between the vectors' norms becomes larger and larger. Thereby, regardless the sign of the cosine index, the greater $\gamma$, the closer to 0 $K^t(\theta, \gamma)$. More formally, we have: $\forall t > 0, \forall\theta$, $\lim_{\gamma \to +\infty} K^t(\theta, \gamma) = 0$.

Notice that for $t < 0$, we observe the opposite effect since the sign of the derivative $\frac{\partial K^t}{\partial \gamma}$ is the same as $\cos\theta$. Thus, in that case, when $\cos\theta > 0$ for example, $K^t(\theta, \gamma)$ is monotonically increasing with respect to $\gamma$. This is not a suitable behavior for a similarity measure, that's the reason why we define $K^t(\theta, \gamma)$ for $t > 0$ only.

In more general terms, the sign and the value of $t$ respectively express the nature and the degree of the influence of $\gamma$ on $K^t(\theta, \gamma)$. First, when $t$ is negative, it defines a coefficient which is not appropriate for measuring similarity. On

[1] Or a monotonically decreasing function with respect to $\theta$ as $\cos\theta$ is a monotonically decreasing function of $\theta$.

**Fig. 1.** Curves respectively representing different values of $K^t(-1,\gamma)$ and $K^t(1,\gamma)$ for different $t$




the contrary, when $t$ is positive it allows one to refine the cosine measure in an appropriate way. Second, assuming that $t > 0$, when the latter increases, it makes $\gamma$ have a more and more important impact on $K^t(\theta,\gamma)$. In that context, it is worthwhile to mention the two following limit cases: when $t \to 0$ we obtain the cosine index which is independent of $\gamma$, whereas when $t \to +\infty$ we have the following index[2]:

$$K^{+\infty}(\mathbf{x},\mathbf{y}) = \frac{\cos\theta}{\max\left(\frac{\|\phi(\mathbf{x})\|}{\|\phi(\mathbf{y})\|}, \frac{\|\phi(\mathbf{y})\|}{\|\phi(\mathbf{x})\|}\right)} = \frac{\cos\theta}{\gamma} \tag{11}$$

To illustrate these points, we plotted in Fig. 1, the graphs corresponding to $K^t(\theta,\gamma)$ for different values of $t$ namely the limit when $t \to 0$, $t = 1$, $t = 10$, $t = 100$ and the limit when $t \to +\infty$. In the left-hand side graph, we fixed $\cos\theta = -1$ and in the right-hand side graph, $\cos\theta$ is set to 1. While the cosine measure is fixed, $\gamma$ varies from 1 to 2 (the horizontal axis). The goal of these graphs is to represent the effects of the norm ratio parameter $\gamma$ on the normalized kernel value (the vertical axis) for different values of $t$ and depending on the sign of the cosine measure. For example, when $\cos\theta = 1$ and $\gamma = 2$, we can observe that, in comparison with the case $t \to 0$ for which the normalized kernel gives 1, the value drops to 0.85 when $t = 1$ and it drops further to 0.5 when $t \to +\infty$.

## 3 Properties of normalized kernels as similarity indices

In this section, we want to better characterize the family of normalized kernels that we have introduced. To this end, we give some relevant properties that they respect. First, we show that $K^t(\mathbf{x},\mathbf{y})$ with $t > 0$ satisfies the basic axioms of

[2] Since we have, $\lim_{t\to+\infty} \mathcal{M}^t(a_1,\ldots,a_p) = \max(a_1,\ldots,a_p)$

geometrical based similarity indices. Second, we show that the normalized kernels of order $t > 0$ are kernels. This result is the main theoretical contribution of this paper.

### 3.1 Basic properties of $K^t$

We start by giving some basic properties of $K^t(\mathbf{x}, \mathbf{y})$ with respect to the general definition of similarity indices defined in a metric space (see for example [6]). $\forall(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^2$, we have:

1 $\forall t > 0 : |K^t(\mathbf{x}, \mathbf{y})| \leq 1$
2 $\forall t > 0 : K^t(\mathbf{x}, \mathbf{x}) = 1$
3 $\forall t > 0 : K^t(\mathbf{x}, \mathbf{y}) = K^t(\mathbf{y}, \mathbf{x})$
4 $\forall t > 0 : \mathbf{x} = \mathbf{y} \Leftrightarrow K^t(\mathbf{x}, \mathbf{y}) = 1$

According to property 1, $K^t(\mathbf{x}, \mathbf{y})$ is bounded by $-1$ and 1 which is a common axiom required for a similarity index. Notice that unnormalized kernels do not respect this axiom in general.

Properties 2 and 3 respectively state that the normalized kernel of order $t$ respects the maximal self-similarity axiom[3] and the symmetric axiom.

According to property 4, the situation $K^t(\mathbf{x}, \mathbf{y}) = 1$ corresponds to the case where the vectors are strictly identical in the feature space. Indeed, considering (9), we can see that the normalized kernel value is 1 if and only if $\cos\theta = 1$ and $\gamma = 1$ which is the same as $\phi(\mathbf{x}) = \phi(\mathbf{y})$. Note that property 4 is not true for the limit case $t \to 0$ for which we only have $\phi(\mathbf{x}) = \phi(\mathbf{y}) \Rightarrow K^0(\mathbf{x}, \mathbf{y}) = 1$. Since for this case the norm ratio plays no role, it is only sufficient for the vectors to be positively colinear to obtain the maximal similarity value 1. Once again, this shows that the normalized kernels of order $t > 0$ are similarity measures that are more discriminative than the cosine measure. It is also worth mentioning in that context, the two other particular cases: both vectors are completely opposite to each other when we observe $K^t(\mathbf{x}, \mathbf{y}) = -1$ ($\cos\theta = -1$ and $\gamma = 1$) and they are geometrically orthogonal when $K^t(\mathbf{x}, \mathbf{y}) = 0$ ($\cos\theta = 0$).

In what follows, we focus on the different relationships between normalized kernels of two distinct orders $t$ and $t'$.

5 $\forall t \geq t' > 0 : \text{sign}(K^t(\mathbf{x}, \mathbf{y})) = \text{sign}(K^{t'}(\mathbf{x}, \mathbf{y}))$
6 $\forall t \geq t' > 0 : \begin{cases} K^t(\mathbf{x}, \mathbf{y}) \leq K^{t'}(\mathbf{x}, \mathbf{y}) & \text{if } \cos\theta > 0 \\ K^t(\mathbf{x}, \mathbf{y}) \geq K^{t'}(\mathbf{x}, \mathbf{y}) & \text{if } \cos\theta < 0 \end{cases}$
7 $\forall t \geq t' > 0 : |K^t(\mathbf{x}, \mathbf{y})| \leq |K^{t'}(\mathbf{x}, \mathbf{y})|$

Property 5 states that the sign of the similarity measure is independent of $t$. More precisely, the sign is only dependent on the angular measure. This is a consequence of (9) since $\mathcal{M}^t(\gamma, \gamma^{-1})$ is strictly positive.

Properties 6 and 7 must be put into relation with the comments we made in section 2 and Fig. 1. Accordingly, these properties formally claim that as $t$

[3] Since property 1 states that 1 is the maximal similarity value. Note that this axiom is also called minimality when dealing with dissimilarity rather than similarity [6].

grows, $K^t(\mathbf{x}, \mathbf{y})$ makes the cosine measure less and less "reliable". We previously mentioned that, all other things being equal, the greater the difference between the vectors' norms in the feature space, the closer to 0 the normalized kernel value. These properties express the fact that this convergence is faster and faster as $t$ grows.

These aforementioned properties are direct consequences of the following relations between generalized means, $\forall 0 < t' \leq t < \infty$:

$$\frac{1}{(\prod_{i=1}^{p} a_i)^{1/p}} > \frac{1}{\mathcal{M}^{t'}(\{a_i\}_{i=1}^{p})} \geq \frac{1}{\mathcal{M}^{t}(\{a_i\}_{i=1}^{p})} > \frac{1}{\max(\{a_i\}_{i=1}^{p})}$$

To complete the analysis of the basic properties that $K^t(\mathbf{x}, \mathbf{y})$ respects with regards to the general axioms required for a similarity index, we need to better characterize the metric properties of the latter. We address this issue in the following subsection.

### 3.2 Metric properties of $K^t$

In this paragraph we denote $K^t$ the similarity matrix of objects in $\mathcal{X}$.

**Theorem 2.** *The similarity matrix $K^t$ with $t > 0$ and general term given by (6) is positive semi-definite. In other words, any normalized kernel $K^t$ with $t > 0$ is a positive semi-definite kernel.*

*Proof (Proof of Theorem 2).*
The proof of this result is based on Gershgorin circle theorem. Let $A$ be a $(n \times n)$ complex matrix with general term $A_{ii'}$. For each row $i = 1, \ldots, n$, its associated Gershgorin disk denoted $\mathcal{D}_i$ is defined in the complex plane as follows:

$$\mathcal{D}_i = \{z \in \mathbb{C} : |z - A_{ii}| \leq \underbrace{\sum_{i' \neq i} |A_{ii'}|}_{R_i}\} = \mathcal{D}(A_{ii}, R_i)$$

Given this definition, Gershgorin circle theorem (see for example [7]) states that all eigenvalues of $A$ lies within $\bigcup_i \mathcal{D}_i$. Accordingly, if the norms of off-diagonal elements of $A$ are small enough then the eigenvalues are not "far" from the diagonal elements. In other words, the lower the $R_i$ quantities, the closer to the diagonal elements the eigenvalues.
Now, let us consider normalized kernel matrices of order $t > 0$ of objects of $\{\mathbf{x}_i; i = 1, \ldots, n\}$. Let denote $K^t_{ii'} = K^t(\mathbf{x}_i, \mathbf{x}_{i'})$. We first suppose the limit case $t' \to 0$. We know that $K^0$ is the cosine similarity matrix. As a consequence, $K^0$ is positive semi-definite and its eigenvalues are all non negative. Next, let us denote $R^t_i = \sum_{i' \neq i} |K^t_{ii'}|$. Then according to properties 2 and 7 given in subsection 3.1, we have:

- $\forall t > 0$ and $\forall i = 1, \ldots, n : K^t_{ii} = 1$,
- $\forall t > t' > 0$ and $\forall i = 1, \ldots, n : R^t_i \leq R^{t'}_i$.

Therefore, when $t$ grows, (6) defines a continuous and differentiable operator for which the absolute value of off-diagonal elements of $K^t$, and consequently the quantities $R_i^t; i = 1, \ldots, n$, are lower and lower while the diagonal entries of $K^t$ remain equal to 1. Thus, applying Gershgorin theorem, we can see that when $t$ increases, the spectrum of $K^t$ is closer and closer to the vector of ones with dimension $n$. As a consequence, since the eigenvalues of $K^0$ are non negative then so are the eigenvalues of $K^t$ with $t > t' > 0$ as the latter are closer to 1 than the former. Finally, for $t > 0$, $K^t$ is symmetric and has non negative eigenvalues. These properties are equivalent to the conditions mentioned in Definition 1 thus, for $t > 0$, we can conclude that $K^t$ are positive semi-definite kernels. □

Finally, a corollary of Theorem 2 [8], is that the related distance $D^t(\mathbf{x}, \mathbf{y}) = \sqrt{2(1 - K^t(\mathbf{x}, \mathbf{y}))}$ respects the triangle inequality axiom, $\forall(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{X}^3$:

8 $\forall t > 0 : D^t(\mathbf{x}, \mathbf{y}) \leq D^t(\mathbf{x}, \mathbf{z}) + D^t(\mathbf{z}, \mathbf{y})$

## 4 Applications to clustering tasks

In order to illustrate the potential benefits of our proposal, we tested different normalized kernels of order $t > 0$ in the context of clustering tasks. The clustering algorithm we used is based on kernel Principal Component Analysis (kernel PCA) and the $k$-means algorithm. Our experiments concern 5 real-world datasets from the UCI ML repository [5]. Our purpose is to show that the normalized kernels that we have introduced, can better capture the proximity relationships between objects compared with other state-of-the-art normalized kernels.

### 4.1 Kernel PCA based $k$-means clustering

Our clustering approach is a spectral clustering like algorithm (see for example [9]). First, from a kernel matrix $K^t$, we proceed to its eigen-decomposition in order to extract from the implicit high dimensional feature space $\mathcal{F}$, an explicit and proper low dimensional representation of the data. Then, we apply a $k$-means algorithm in that reduced space in order to find a partition of the objects.

Principal Component Analysis (PCA) is a powerful and widely used statistical technique for dimension reduction and features extraction. This technique was extended to kernels in [3]. Formally, let denote $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{k-1}$ the leading $k-1$ eigenvalues of $K^t$ and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{k-1}$ the corresponding eigenvectors. The low dimensional space extracted from $K^t$ that we used as data representation for the clustering step is spanned according to:
$(\sqrt{\lambda_1}\mathbf{v}_1, \sqrt{\lambda_2}\mathbf{v}_2, \ldots, \sqrt{\lambda_{k-1}}\mathbf{v}_{k-1})$.

When applying dimension reduction techniques prior to a clustering algorithm, an important issue is the number of dimensions that one has to retain. In this paper, since we aim at using the $k$-means algorithm as the clustering method, we follow the work presented in [4] that concerns the relationship between $k$-means and PCA. Accordingly, if $k$ is the number of clusters to find then we retain the $k-1$ leading eigenvectors.

With regards to related works, we can cite the following papers that use Kernel PCA based clustering in the contexts of image and text analysis respectively, [10, 11].

### 4.2 Experiments settings

The datasets that we used in our experiments are the following ones [5]:

- Iris (150 objects, 4 features in the input space, 3 clusters)
- Ecoli (336 objects, 7 features in the input space, 8 clusters)
- Pima Indian Diabetes (768 objects, 8 features in the input space, 2 clusters)
- Yeast (1484 objects, 8 features in the input space, 8 clusters)
- Image Segmentation (2310 objects, 18 features in the input space, 7 clusters)

For each data set, we first normalized the data in the input space by centering and standardizing the features. Next, we applied different kernels $K$ namely, the linear kernel $K_l(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle$ which is simply the dot product in the input space, and the polynomial kernel $K_p(\mathbf{x}, \mathbf{y}) = (\langle \mathbf{x}, \mathbf{y} \rangle + 1)^2$. For each type of kernels, we computed different normalized kernels $K^t$ by varying the value of $t$: $t \to 0$, $t = 1$, $t = 10$, and the limit $t \to +\infty$. To each of those similarity matrices, we applied the kernel PCA clustering method described previously. Since the $k$-means algorithm is sensitive with respect to the initialization, for all cases we launched the algorithm 5 times with different random seeds and took the mean average value of the assessment measures.

Since we deal with normalized kernels that amount to projecting the objects on an unit hypersphere, we also tested the kernel PCA based $k$-means clustering with the Gaussian RBF kernel which presents the same property. This case is our first baseline. However, when using such a kernel, one has to tune a parameter $\sigma$. In this paper, to get rid of this problem, we applied the approach proposed in [12] which suggests to use the following affinity measure:

$$K_g(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{y}\|^2}{\sigma_{\mathbf{x}} \sigma_{\mathbf{y}}}\right) \tag{12}$$

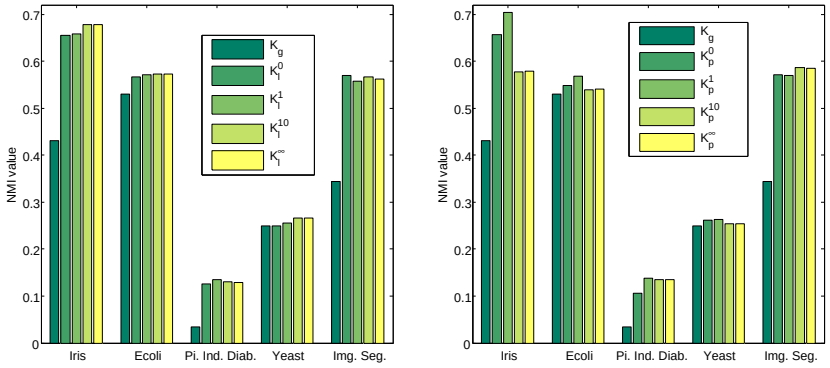
where $\sigma_{\mathbf{x}}$ is set to the value of the distance between $\mathbf{x}$ and its 7th nearest neighbor.

Besides, as our purpose is to show that taking into account the mapped vectors' norm ratio in addition to the cosine measure can be beneficial, we took the case $t \to 0$, which simply corresponds to the cosine index, as a baseline as well.

The assessment measure of the clustering outputs we used is the Normalized Mutual Information (NMI) introduced in [13]. Let denote $U$ the partition found by the kernel PCA clustering and $V$ the ground-truth partition. This evaluation measure is denoted $\mathrm{NMI}(U, V)$ and is given by:

$$\mathrm{NMI}(U, V) = \frac{\sum_{u=1}^{k} \sum_{v=1}^{k} N_{uv} \log\left(\frac{n N_{uv}}{N_{u.} N_{.v}}\right)}{\sqrt{\left(\sum_{u=1}^{k} N_{u.} \log\left(\frac{N_{u.}}{n}\right)\right)\left(\sum_{v=1}^{k} N_{.v} \log\left(\frac{N_{.v}}{n}\right)\right)}} \tag{13}$$

**Fig. 2.** NMI measures for the different datasets with different kernels ($K_l$ on the left and $K_p$ on the right) and different normalizations ($t \to 0$,$t = 1$,$t = 10$,$t \to +\infty$)




where $N$ is the contingency table between $U$ and $V$ and $n$ the total number of objects. $\text{NMI}(U, V)$ lies in $[0, 1]$ and the higher the measure, the better the clustering output $U$.

### 4.3 Experiments results

In Fig. 2, we report the results we obtained for each dataset. On the left-hand side we give the results related to the linear kernel whereas on the right-hand side, the NMI values correspond to the polynomial kernel. In both graphs, the results provided by the Gaussian RBF kernel are shown (1st bar). Compared to this baseline we can see that the cosine measure and the normalized kernels all perform better on the datasets used in these experiments.

In comparison with the other baseline $K^0$, we can observe that in most cases there are normalized kernels of order $t > 0$ which can lead to better NMI values. When using the linear kernel, this is true for all $K^t$ except for the Image Segmentation dataset. Furthermore, for the Iris, Ecoli and Yeast datasets, as $t$ grows the performances are consistently better. This shows that taking into account the vectors' norm ratio in the feature space in order to refine the cosine measure, is beneficial.

In the case of polynomial kernel, not all normalization techniques are interesting since many normalized kernels do not outperform the cosine measure. However, when using this kernel, it seems that taking $K_p^1$ as a normalization method is a good choice. Particularly, $K_p^1$ leads to the best performances for the Iris and the Pima Indian Diabetes datasets. Besides, concerning the Image Segmentation dataset, we can see that unlike the linear kernel, the normalized polynomial kernels can outperform the cosine index since the best result is obtained with $K_p^{10}$.

## 5 Conclusion and future work

We have introduced a new family of normalization methods for kernels which extend the cosine normalization. We have detailed the different properties of such methods and the resulting proximity measures with respect to the basic axioms of geometrical based similarity indices. Accordingly, we have shown that normalized kernels are "proper" similarity indices that amount to projecting the data on an unit hypersphere. We have, in addition, proved that these normalized kernels are also kernels. From a practical standpoint, we have also validated the utility of normalized kernels in the context of clustering tasks using several real-world datasets. However, one remaining issue is the choice of the order $t$ when normalizing a kernel. We have shown from a theoretical and a practical point of view that the norm ratio measure can make the normalized kernel more efficient but still, the weight one should give to this parameter in comparison with the angular measure is not straightforward to set. In our future work we intend to further investigate this problem.

## References


1. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA (2001)
2. Vert, J., Tsuda, K., Schlkopf, B.: A primer on kernel methods. In Schlkopf, B., K.T., Vert, J., eds.: Kernel Methods in Computational Biology, Cambridge, MA, USA, MIT Press (2004) 35–70
3. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. **10**(5) (1998) 1299–1319
4. Ding, C., He, X.: K-means clustering via principal component analysis. In: ICML '04: Proceedings of the twenty-first international conference on Machine learning, New York, NY, USA, ACM (2004) 29
5. Asuncion, A., Newman, D.: UCI machine learning repository (2007)
6. Santini, S., Jain, R.: Similarity measures. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(9) (1999) 871–883
7. Horn, R., Johnson, C.: Matrix analysis. Cambridge University Press (1985)
8. Gower, J., Legendre, P.: Metric and euclidean properties of dissimilarity coefficients. Journal of classification **3** (1986) 5–48
9. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. Pattern Recogn. **41**(1) (2008) 176–190
10. Tuytelaars, T., Lampert, C.H., Blaschko, M.B., Buntine, W.: Unsupervised object discovery: A comparison. International Journal of Computer Vision **Epub ahead** (07 2009) 1–19
11. Minier, Z., Csató, L.: Kernel PCA based clustering for inducing features in text categorization. In: ESANN. (2007) 349–354
12. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: Advances in Neural Information Processing Systems 17. (2005)
13. Strehl, A., Strehl, E., Ghosh, J.: Cluster ensembles - a knowledge reuse framework for combining partitionings. Journal of Machine Learning Research **3** (2002) 583–617