Designing a Metric for the Difference Between Gaussian Densities

Karim T. Abou–Moustafa, Fernando De La Torre and Frank P. Ferrie

Abstract Measuring the difference between two multivariate Gaussians is central to statistics and machine learning. Traditional measures based on the Bhattacharyya coefficient or the symmetric Kullback–Leibler divergence do not satisfy metric properties necessary for many algorithms. This paper proposes a metric for Gaussian densities. Similar to the Bhattacharyya distance and the symmetric Kullback–Leibler divergence, the proposed metric reduces the difference between two Gaussians to the difference between their parameters. Based on the proposed metric we introduce a symmetric and positive semi-definite kernel between Gaussian densities. We illustrate the benefits of the proposed metric in two settings: (1) a supervised problem, where we learn a low-dimensional projection that maximizes the distance between Gaussians, and (2) an unsupervised problem on spectral clustering where the similarity between samples is measured with our proposed kernel. [1]

Karim T. Abou–Moustafa
Centre of Intelligent Machines (CIM), McGill University, 3480 University street, Montreal, QC, H3A 2A7, CANADA, e-mail: karimt@cim.mcgill.ca

Fernando De La Torre
The Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, USA e-mail: ftorre@cs.cmu.edu

Frank P. Ferrie
Centre of Intelligent Machines (CIM), McGill University, 3480 University street, Montreal, QC, H3A 2A7, CANADA, e-mail: ferrie@cim.mcgill.ca

[1] This is the main McGill CIM Technical Report (#TR–CIM–10–05) for our research describing a measure for the difference between Gaussian densities. This technical report was presented at the *International Symposium on Brain, Body and Machine* – Nov. 10–12, 2010 – which was held to celebrate the 25th anniversary of McGill Centre of Intelligent Machines (CIM). The first and third authors are affiliated with CIM. The proceedings of the symposium have been published in this Springer Series on *Advances in Intelligent and Soft Computing.*

1 Introduction

The Gaussian distribution plays a crucial role in multivariate statistics in general, and in discrimination theory in particular [1]. A direct realization of this fact is to note how Gaussian densities are pervasive in statistical machine learning. A major aspect in discrimination theory, and consequently in statistical learning, is to reflect how two probability distributions are *close to*, or *far away* from each other; or more formally, quantify the separation or similarity/dissimilarity between probability distributions. Recently, there has been an increasing interest in defining dissimilarity measures on probability distributions to tackle problems involving structured data and/or objects not in vectorized form, when locally represented by generative models or probability distributions [11, 26, 17, 20, 12, 19]. If $\mathcal{X} = \{X_1, \ldots, X_n\}$ is the input space of such data points or objects (images, documents, proteins, variable length sequences of audio or video frames, etc.), and $\mathcal{P}$ is the space of a certain parametric family of probability distributions, then handling this type of data is usually done by mapping each datum from $\mathcal{X}$ to a probability distribution in $\mathcal{P}$. Hence, defining a dissimilarity measure on $\mathcal{P}$ in fact induces a dissimilarity measure on $\mathcal{X}$.

Our contribution in this paper is three–fold. Due to the importance of the Gaussian distribution, we define a separation or dissimilarity measure for the family of d-dimensional Gaussian distributions $\mathcal{G}_d$, such that the measure, among other requirements, should be a *full metric*; i.e. satisfy the three metric axioms: positivity, symmetry and obey the triangle inequality. Based on the three metric axioms satisfied by our metric, (1) we propose a kernel between Gaussian densities and show that it is symmetric and positive semi–definite (PSD), and (2) define an embedding for the objects in $\mathcal{X}$ into a low dimensional subspace $\mathbb{R}^{d_0}$ where $d_0 \ll n$. As it will be shown here, (1) and (2) can not be achieved if any of the three metric axioms are not satisfied.

Our proposed measure is in many ways very similar to the closed form expressions of the Bhattacharyya divergence [3] and the symmetric Kullback–Leibler (KL) divergence [16] between two multivariate Gaussian densities. However, unlike those measures of divergence that are positive, symmetric, and violate the triangle inequality [13], our proposed metric meets the three metric axioms. As will be discussed below, all measures of divergence for probability distributions are positive by definition of the divergence and can be symmetrized [1, 6]. However, very few of them meet the triangle inequality axiom.

Since our proposed measure is a full metric (by definition) on $\mathcal{G}_d$, then mapping from $\mathcal{X}$ to $\mathcal{G}_d$ yields interesting consequences for various learning algorithms. **First**; most classification and clustering algorithms assume that $\mathcal{X} \subseteq \mathbb{R}^d$ and hence, they rely on the Euclidean measure to define distances/similarities between points. If objects in $\mathcal{X}$ are complex structured data – variable length time series data or not in vectorized form – it becomes very difficult to apply these algorithms on such data. However, mapping these objects from $\mathcal{X}$ to $\mathcal{G}_d$ and using our proposed metric alleviates this difficulty by using these algorithms on their images in $\mathcal{G}_d$. **Second**; there have been some serious advances recently in speeding up the k–means algorithm by avoiding many distance computations between points and cluster centres [8]. This was possible to achieve through the triangle inequality property of the Euclidean metric to compute upper and lower bounds on these distances. It is straight forward that our proposed metric can use these same bounds to speed up clustering in $\mathcal{G}_d$. **Third**; by exponentiating the negative value of the metric, one directly obtains a kernel $K_{\mathcal{G}} : \mathcal{G}_d \times \mathcal{G}_d \to \mathbb{R}$ that, as will be shown here, is symmetric and PSD [5, 10]. This allows a smooth extension for all kernel based methods [23] to be applied on objects mapped to $\mathcal{G}_d$.

The triangle inequality axiom, in addition, allows us to consider a more general aspect of our proposed measure. If $\mathbf{D}_{\mathcal{G}} \in \mathbb{R}^{n\times n}$ is a symmetric matrix with zero diagonal elements (self distances) and filled with the mutual distances between the n objects in $\mathcal{X}$ using our proposed metric on $\mathcal{G}_d$, and $\tilde{\mathbf{D}}_{\mathcal{G}}$ is the centralized[2] distance matrix of $\mathbf{D}_{\mathcal{G}}$, then : (1) $\mathbf{G} = -\frac{1}{2}\tilde{\mathbf{D}}_{\mathcal{G}}$ is a PSD matrix that defines a dot product (or a gram) matrix in a Hilbert space, (2) there exists a matrix $\mathbf{X} \in \mathbb{R}^{n\times d_0}$ s.t. $\mathbf{G} = \mathbf{X}\mathbf{X}^\top$ that provides for the objects in $\mathcal{X}$ an embedding in $\mathbb{R}^{d_0}$, and the dimensionality d_0 is the rank of the matrix $\mathbf{G}$, and (3) for the case of $n = 3$, that $\mathbf{G}$ is PSD is equivalent to the triangular inequality relation between the three points. These results are credited to Young and Householder [27] (and recently by Roth *et al.* [22]) who establish the equivalence between the triangle inequality axiom of a metric and the positive semi–definiteness of the gram matrix $\mathbf{G}$.

Our research work starts by analyzing the closed form expressions for the Bhattacharyya divergence and the symmetric KL divergence between two multivariate Gaussian densities. We note that both have very similar properties and structure with regard to their closed form expression. Next, we propose our dissimilarity metric for Gaussian densities. Using this proposed metric, we introduce a kernel for Gaussian densities and show that it is symmetric and PSD. Finally, using preliminary experiments, we validate the proposed metric in two settings; (1) supervised, where the metric is maximized to learn a lower dimensional subspace for discriminant analysis, and (2) unsupervised, where our proposed kernel is used with spectral clustering to measure the similarity between images.

2 Related work

Earlier work on dissimilarity measures for probability distributions started with kernels for generative models in order to plug them in discriminative

[2] $\tilde{\mathbf{D}}_{\mathcal{G}} = \mathbf{Q}\mathbf{D}_{\mathcal{G}}\mathbf{Q}$, where $\mathbf{Q} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$; i.e. the row sum and column sum of $\tilde{\mathbf{D}}_{\mathcal{G}}$ is zero.

models such as SVMs. This includes the Fisher kernel by Jaakkola and Haussler [11] and then the TOP kernel by Tsuda *et al.* [26].

Lafferty and Lebanon [17] study the statistical manifold of probability distributions and arrive at heat or diffusion kernels. In particular, they find closed form expressions for Gaussian distributions with spherical covariances and for multinomial distributions, where the latter is used for text documents classification. Moreno *et al.* [20] exponentiate the negative symmetric KL divergence (2) to define a kernel between Gaussian distributions for an SVM classifier: $K_{KL} = \exp\{-\alpha d_{KL} + \beta\}$, where α and β are scaling and shifting parameters respectively. Since K_{KL} is not PSD, the authors raise the need for α and β to scale and shift the kernel until it gets closer to a PSD one. Jebara and Kondor [12] define the probability product kernel (PPK) as a general symmetric and PSD kernel that can be applied to various probability distributions and generative models: $K(P_1, P_2) = \int_{\mathcal{X}} p_1^\alpha(\mathbf{x}) p_2^\alpha(\mathbf{x}) = \langle p_1^\alpha, p_2^\alpha \rangle_{L_2}$, where α is a positive constant. They consider two cases for α; 1) $\alpha = 1/2$, where they arrive at the Bhattacharyya affinity ρ, and 2) $\alpha = 1$, where they arrive at the expected likelihood kernel. More recently, Martins *et al.* [19] expand the set of kernels based on information theoretic measures by incorporating nonextensive information measures as similarity measures between probability distributions.

In independent and different research paths, Roth *et al.* [22] develop a formal treatment for correcting dissimilarity measures that do not satisfy the triangle inequality based on the results of Young and Householder [27]. The kernel community, in another direction, has recently developed a framework for distances between probability distributions based on a Hilbert space embedding of these distributions without explicit density estimation. They further plug this distance it in a variety of problems arising in statistics such as homogeneity tests, independence measurement and feature selection. Please refer to Sriperumbudur *et al.* [25] for recent advances and results in this direction.

3 Divergences and distances for probability distributions

In statistics and information theory, dissimilarity measures of probability distributions are known as coefficients of divergence, Ali–Silvey distances [1], or f–divergence according to Csiszar [6]. If $P_1, P_2 \in \mathcal{P}$ are two probability distributions defined over the same domain of events $\mathcal{E}$, then the divergence of P_2 from P_1 is defined as $d_f(P_1, P_2) = \mathbb{E}_{p_1}\{C(\phi)\} = \int_{\mathcal{E}} p_1(x) C(\phi(x))\, dx$, where $d_f(P_1, P_2) \in [0, \infty)$, 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[3], and C is a continuous convex function on $(0, \infty)$.

[3] The original definition of ϕ is the generalized Radon–Nikodym derivative of P_1 with respect to P_2

The divergence, according to Ali & Silvey or Csiszar, has to satisfy certain requirements. The most relevant to our discussion is that it should be zero when $P_1 = P_2$ and as large as possible when P_1 and P_2 are farthest apart. This is exactly the first axiom of a metric. However, the divergence by definition, is not symmetric and need not to obey the triangle inequality. Although, any divergence can be transformed to a symmetrized measure by summing $d_f(P_1, P_2)$ and $d_f(P_2, P_1)$, it is neither trivial nor obvious how to satisfy the triangle inequality. For the purpose of our discussion, we shall consider the symmetric KL divergence [16] and Chernoff's measure for discriminatory information [4] which yields the Bhattacharyya coefficient [3] and consequently the Bhattacharyya divergence [3] and the Hellinger distance [21]. Note that all these measures can be directly derived from $d_f(P_1, P_2)$ (see [1] for more details).

3.1 Distances and divergences for Gaussian densities

Before proceeding, we need to introduce our notation for multivariate Gaussian densities. Let $\{\mathcal{N}_j(\mathbf{x};\mu_\mathbf{j},\mathbf{\Sigma_j}) \in \mathcal{G}_\mathbf{d} \mid \mathbf{x}, \mu \in \mathbb{R}^\mathbf{d},\ \mathbf{\Sigma_j} \in \mathbb{S}_{++}^{\mathbf{d\times d}},\ \mathbf{j = 1, 2}\}$ be two Gaussian densities where:

$$\mathcal{N}_j(\mathbf{x};\mu_\mathbf{j},\mathbf{\Sigma_j}) = |2\pi\mathbf{\Sigma}_j|^{-\frac{1}{2}} \exp\{-\tfrac{1}{2}(\mathbf{x}-\mu_\mathbf{j})^\top \mathbf{\Sigma_j^{-1}}(\mathbf{x}-\mu_\mathbf{j})\}, \tag{1}$$

$|\cdot|$ is the determinant, $\mu_\mathbf{j}$ is the mean vector, $\mathbf{\Sigma_j}$ is the covariance matrix, and $\mathbb{S}_{++}^{d\times d}$ is the space of real symmetric PSD matrices.

The symmetric KL divergence is based on Kullback's measure of discriminatory information: $I(P_1, P_2) = -\int_\mathcal{E} p_1 \log(p_1/p_2)dx$. Kullback realizes the asymmetry of $I(P_1, P_2)$ and describes it as the *directed divergence*. To achieve symmetry, Kullback defines the divergence as $I(P_1, P_2)+I(P_2, P_1)$ and notes that it is positive and symmetric but violates the triangle inequality [16] (p. 6,7). Hence, it can not define a metric structure. The closed form expression for the symmetric KL divergence between $\mathcal{N}_1$ and $\mathcal{N}_2$ can be written as:

$$d_{KL}(\mathcal{N}_1,\mathcal{N}_2) = \tfrac{1}{2}\mathbf{u}^\top(\mathbf{\Sigma}_1^{-1}+\mathbf{\Sigma}_2^{-1})\mathbf{u} + \tfrac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2+\mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1-2\mathbf{I}), \tag{2}$$

where tr is the matrix trace, $\mathbf{u} = (\mu_\mathbf{1} - \mu_\mathbf{2})$, and $\mathbf{I}$ is the identity matrix. Equation (2) describes d_{KL} as a sum of two components, one due to the difference in means weighted by the covariance matrices, and the other due to the difference in variances and covariances [16] (p. 6,7). If $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, then d_{KL} expresses the difference in means which is the exact form of the Mahalanobis distance: $(\mu_\mathbf{1} - \mu_\mathbf{2})^\top\mathbf{\Sigma^{-1}}(\mu_\mathbf{1} - \mu_\mathbf{2})$. However, if $\mu_\mathbf{1} = \mu_\mathbf{2} = \mu$, then d_{KL} expresses the difference, or the dissimilarity between covariance matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$:

$$d_{KL}(\mathcal{N}_1, \mathcal{N}_2) = \tfrac{1}{2}\mathrm{tr}(\mathbf{\Sigma}_1^{-1}\mathbf{\Sigma}_2 + \mathbf{\Sigma}_2^{-1}\mathbf{\Sigma}_1 - 2\mathbf{I}). \tag{3}$$

The Bhattacharyya divergence, on the other hand, is a special case of Chernoff's [4] measure of discriminatory information: $d_{Ch}(P_1, P_2) = -\ln(\inf_{0<t<1} \int_{\mathcal{E}} p_1^t p_2^{1-t} dx)$. Setting $t = 1/2$, although not the infimum but still within the valid range, yields the Bhattacharyya divergence [3]: $d_B(P_1, P_2) = -\ln \int_{\mathcal{E}} \sqrt{p_1 p_2} dx = -\ln \rho(P_1, P_2)$, where ρ is the Bhattacharyya coefficient. Note that $0 \leq d_B \leq \infty$ and $0 \leq \rho \leq 1$. The coefficient ρ can define another distance: $d_H(P_1, P_2) = \sqrt{1 - \rho(P_1, P_2)}$, $0 \leq d_H \leq 1$, which is known as the Hellinger distance [21]. Kailath [13] carefully studied d_B and d_H and notes that d_B is positive and symmetric but violates the triangle inequality, while d_H meets all axioms that define a metric. Here, we also note the work of Jebara and Kondor [12] who arrive to the Bhatacharyya coefficient ρ via the probability product kernel (PPK). They define ρ as the Bhattacharyya affinity and confirm through the PPK definition that ρ is a PSD kernel.

The closed form for the Bhattacharyya coefficient ρ between $\mathcal{N}_1$ and $\mathcal{N}_2$ can be written as follows:

$$\rho(\mathcal{N}_1, \mathcal{N}_2) = \frac{|\mathbf{\Sigma}_1|^{\frac{1}{4}}|\mathbf{\Sigma}_2|^{\frac{1}{4}}}{|\frac{1}{2}\mathbf{\Sigma_1} + \frac{1}{2}\mathbf{\Sigma}_2|^{\frac{1}{2}}} \exp\{-\tfrac{1}{8}\mathbf{u}^\top(\tfrac{1}{2}\mathbf{\Sigma}_1 + \tfrac{1}{2}\mathbf{\Sigma}_2)^{-1}\mathbf{u}\}. \tag{4}$$

The closed form of the Hellinger distance between $\mathcal{N}_1$ and $\mathcal{N}_2$ is directly obtained from the Bhattacharyya coefficient, however the expression for the Bhattacharyya divergence has a more interesting compact form:

$$d_B(\mathcal{N}_1, \mathcal{N}_2) = \tfrac{1}{8}\mathbf{u}^\top(\tfrac{1}{2}\mathbf{\Sigma}_1 + \tfrac{1}{2}\mathbf{\Sigma}_2)^{-1}\mathbf{u} + \tfrac{1}{2}\ln\frac{\frac{1}{2}|\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|^{\frac{1}{2}}|\mathbf{\Sigma}_2|^{\frac{1}{2}}}. \tag{5}$$

Similar to d_{KL} in Equation (2), d_B in Equation (5) is expressed as the sum of two components, one due to the difference in means, and the other due to the difference in covariance matrices. If $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, then d_B is equal to the Mahalanobis distance up to a scaling factor $(\frac{1}{8})$, and if $\mu_{\mathbf{1}} = \mu_{\mathbf{2}} = \mu$, then d_B will express the dissimilarity between the matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$:

$$d_B(\mathcal{N}_1, \mathcal{N}_2) = \tfrac{1}{2}\ln\frac{\frac{1}{2}|\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|^{\frac{1}{2}}|\mathbf{\Sigma}_2|^{\frac{1}{2}}}. \tag{6}$$

3.2 A close look at d_{KL} and d_B

We note that when the Bhattacharyya divergence and the symmetric KL divergence were applied to $\mathcal{N}_1$ and $\mathcal{N}_2$, they factored the difference between the distributions in terms of the difference between their first and second order statistics. In other words, the difference between two Gaussian densities was reduced to the difference between their parameters. Note also that d_{KL} and

d_B in Equations (2) and (5) respectively have the same structure; a sum of two components, one due to the difference in means (represented as a Mahalanobis distance), and the other due to the difference in covariance matrices. More precisely, the first component in d_{KL} is the sum of two Mahalanobis distances, while the first component in d_B is a variant of the Mahalanobis distance that uses the inverse of an average covariance matrix. Note that this explanation for the meaning of each term is due to Kullback [16] (p. 6,7). The Mahalanobis distance comprising the first component of d_{KL} and d_B meets the three metric axioms. However, since Equations (2) & (5) are positive and symmetric but violate the triangle inequality, then the reason for the deficiency in meeting the triangle inequality is due to the second component in d_{KL} and d_B which measures the dissimilarity between covariance matrices, i.e. Equations (3) and (6). This observation implies that the measures for the difference between the PSD covariance matrices in Equations (3) and (6) do not define proper metrics for covariance matrices on the manifold $\mathbb{S}_{++}^{d\times d}$.

4 Designing a metric for Gaussian densities

The discussion above suggests that if there is a distance measure for covariance matrices that defines a metric on the manifold $\mathbb{S}_{++}^{d\times d}$, then it is possible to design a new separation measure specifically for Gaussian densities. The designed measure however, should meet certain requirements: 1) it should satisfy all the metric axioms, 2) reduce to the Euclidean distance when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}$, 3) reduce to the Mahalanobis distance when $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{\Sigma}$, and 4) reduce to a metric for covariance matrices, satisfying all metric axioms, when $\mu_{\mathbf{1}} = \mu_{\mathbf{2}} = \mu$.

Note that requirements 2 & 3 are in the same spirit of d_{KL} and d_B in Equations (2) and (5) which provide an intuitive meaning and explanation for the metric in any of these special cases. Although the Hellinger distance d_H is a true metric by definition, it does not yield such properties, however d_H has a defined range ($d_H \in [0, 1]$) which might be desirable in certain contexts.

Similar to d_{KL} and d_B, the new designed metric will comprise two components, a measure for the difference in means, and a measure for the difference in covariance matrices. However, unlike these measures, the second component will be a true metric for real symmetric PSD matrices on the manifold $\mathbb{S}_{++}^{d\times d}$.

4.1 A metric for symmetric and positive semi–definite matrices

Förstner and Moonen [9] proposed a metric measure for covariance matrices that is derived from a canonical invariant Riemannian metric on the manifold $\mathbb{S}_{++}^{d\times d}$. The measure is a full metric, invariant under affine transformations of the coordinate system, and invariant to inversion. For two matrices $\{\mathbf{A}, \mathbf{B} \in \mathbb{S}_{++}^{d\times d}\}$ the distance measure between them is:

$$d_{FM}(\mathbf{A}, \mathbf{B}) = \sqrt{\mathrm{tr}\{\ln^2 \mathbf{\Lambda}(\mathbf{A}, \mathbf{B})\}}, \tag{7}$$

where $\mathbf{\Lambda}(\mathbf{A}, \mathbf{B}) = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ is the solution of a generalized eigenvalue problem (GEP): $\mathbf{AV} = \mathbf{\Lambda BV}$. The proof that d_{FM} defines a metric on the manifold $\mathbb{S}_{++}^{d\times d}$ and that it satisfies all the axioms of a metric can be found in [9]. The basic idea of comparing covariance matrices is to reflect the deviations in variances in all directions. In d_{FM}, these deviations are evaluated as the ratio of variances for all dimensions. The ln in d_{FM} measures these deviations as factors, while squaring guarantees that deviations by a factor of f and $1/f$ will be equally penalized.

4.2 The proposed metric $d_{\mathcal{G}}$

Our metric is designed based on the first component of the Bhattacharyya distance for the difference in means, and on the metric d_{FM} in Equation (7) for covariance matrices. For two Gaussian densities $\mathcal{N}_1$ and $\mathcal{N}_2$, the proposed metric $d_{\mathcal{G}}$ is defined as follows:

$$d_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2) = \left(\mathbf{u}^\top \mathbf{S}^{-1} \mathbf{u}\right)^{\frac{1}{2}} + \left(\sum_{k=1}^{d} \ln^2 \lambda_k(\mathbf{\Sigma}_1, \mathbf{\Sigma}_2)\right)^{\frac{1}{2}}. \tag{8}$$

Except for the invariance to inversion property of d_{FM}, $d_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2)$ inherits all the properties of its constituting components: 1) it is invariant to affine transformations, 2) it is a full metric, and 3) it fulfils the special cases of requirements 2, 3 & 4 mentioned above.

Similar to the Bhattacharyya distance (5) and the symmetric KL divergence (2), the proposed metric $d_{\mathcal{G}}$ reduces the difference between $\mathcal{N}_1$ and $\mathcal{N}_2$ to the difference between their parameters. Moreover, it has the exact same structure as d_{KL} and d_B, where the first term measures the difference in means, while the second term measures the difference between two covariance matrices (i.e. two symmetric and PSD matrices). In other words, each term in $d_{\mathcal{G}}$ has a clear meaning and measures a well defined quantity.

4.3 A kernel based on $d_{\mathcal{G}}$

We can define a kernel $K_{\mathcal{G}} : \mathcal{G}_d \times \mathcal{G}_d \to \mathbb{R}$ for two Gaussian densities based on $d_{\mathcal{G}}$ as follows:

$$K_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2) = \exp\{-d_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2)\}. \tag{9}$$

The kernel $K_{\mathcal{G}}$ is an exponential function of a distance measure that is not an Euclidean norm. Genton in [10] studies different classes of kernels with their properties and points to [5] for a formal treatment for the case of $K_{\mathcal{G}}$. In particular, Christakos and Papanicolaou [5] set conditions for the class of exponential kernels when the distance that defines the kernel is not an Euclidean metric on $\mathbb{R}^d$. To show that $K_{\mathcal{G}}$ is a PSD kernel, we rewrite $d_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2)$ in Equation (8) as follows:

$$\begin{aligned} d_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2) &= (\mathbf{u}^\top \mathbf{\Phi}^\top \mathbf{\Gamma}^{-1} \mathbf{\Phi} \mathbf{u})^{\frac{1}{2}} + \left(\sum_{i=1}^{d} \ln^2 \lambda_i \right)^{\frac{1}{2}} \\ &= \left(\sum_{j=1}^{d} \gamma_j (\mathbf{u}^\top \phi_{\mathbf{j}})^{\mathbf{2}} \right)^{\frac{1}{2}} + \left(\sum_{i=1}^{d} \omega_i \ln^2 \lambda_i \right)^{\frac{1}{2}}, \text{ and hence} \end{aligned}$$

$$K_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2) = \exp\left\{ - \left(\sum_{j=1}^{d} \gamma_j (\mathbf{u}^\top \phi_{\mathbf{j}})^{\mathbf{2}} \right)^{\frac{1}{2}} \right\} \exp\left\{ - \left(\sum_{i=1}^{d} \omega_i \ln^2 \lambda_i \right)^{\frac{1}{2}} \right\}, \tag{10}$$

where $\omega_i = 1$, for $1 \le i \le d$, $\mathbf{\Phi} = [\phi_{\mathbf{1}} \dots \phi_{\mathbf{d}}]$ is a column matrix with the eigenvectors of $\mathbf{S}$, and $\mathbf{\Gamma} = \operatorname{diag}(\gamma_1, \dots, \gamma_d)$ is the diagonal matrix of its eigenvalues. To show that $K_{\mathcal{G}}$ is a PSD kernel, one has to show that each term in the right hand side (RHS) of (10) is a PSD kernel since the multiplication of two PSD kernels is another PSD kernel [10] (p. 300). Let the lag vectors $\mathbf{h}_1$ and $\mathbf{h}_2$ and the weight vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ be respectively defined as follows:

$$\mathbf{h}_1 = [\mathbf{u}^\top \phi_{\mathbf{1}}, \dots, \mathbf{u}^\top \phi_{\mathbf{d}}]^\top, \quad \mathbf{h_2} = [\ln \lambda_{\mathbf{1}}, \dots, \ln \lambda_{\mathbf{d}}]^\top, \tag{11}$$

$$\mathbf{w}_1 = [\gamma_1, \dots, \gamma_d]^\top, \quad \mathbf{w}_2 = [\omega_1, \dots, \omega_d]^\top. \tag{12}$$

It is shown in [5] (p. 475) that if an exponential kernel K is of the form:

$$K(\mathbf{h}) = \exp\{-(w_1 |h_1|^p + \dots + w_n |h_n|^p)^{\frac{1}{2}}\}, \tag{13}$$

for a lag vector $\mathbf{h} \in \mathbb{R}^n$ and a weight vector $\mathbf{w} \in \mathbb{R}^n$, then K is a PSD kernel if and only if $0 < p \le 2$. Setting $p = 2$ and using the definition of $\mathbf{h}_1$, $\mathbf{h}_2$, $\mathbf{w}_1$ and $\mathbf{w}_2$ from (11) and (12), then $K_{\mathcal{G}}$ can be written as:

$$K_{\mathcal{G}}(\mathcal{N}_1,\mathcal{N}_2) = \exp\left\{-\left(w_1^1(h_1^1)^2 + \cdots + w_d^1(h_d^1)^2\right)^{\frac{1}{2}}\right\} * \exp\left\{-\left(w_1^2(h_1^2)^2 + \cdots + w_d^2(h_d^2)^2\right)^{\frac{1}{2}}\right\}, \qquad (14)$$

where each term on the RHS of (14) has the exact same structure of (13), and hence each term defines a PSD kernel. Consequently, it follows that $K_{\mathcal{G}}$ is a PSD kernel.

Note that the definition of $K(\mathbf{h})$ in Equation (13) allows the introduction of a kernel parameter $\sigma > 0$ in $K_{\mathcal{G}}$ that controls the affinity between the Gaussian densities (in fact there parameters) without loosing its PSD property. Therefore, the final form of our proposed kernel is : $\exp\{-d_{\mathcal{G}}(\mathcal{N}_1,\mathcal{N}_2)/\sigma\}$ and $\sigma > 0$.

Finally, if $\mathbf{W}_{\mathcal{G}} \in \mathbb{R}^{n\times n}$ is the kernel or gram matrix obtained from $K_{\mathcal{G}}(\mathcal{N}_i,\mathcal{N}_j)$, for $1 \leq i,j \leq n$, then $\mathbf{W}_{\mathcal{G}}$ meets with the gram matrix $\mathbf{G} = -\frac{1}{2}\tilde{\mathbf{D}}_{\mathcal{G}}$ (defined earlier) in that both are derived from $\mathbf{D}_{\mathcal{G}}$ and both are symmetric and PSD. Therefore, using the results in [27, 22], there exist matrices $\mathbf{X}_1 \in \mathbb{R}^{n\times d_1}$ and $\mathbf{X}_2 \in \mathbb{R}^{n\times d_2}$ s.t. $\mathbf{G} = \mathbf{X}_1\mathbf{X}_1^\top$ and $\mathbf{W}_{\mathcal{G}} = \mathbf{X}_2\mathbf{X}_2^\top$ that provide an embedding for the objects in $\mathcal{X}$ into a lower dimensional space $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$, and the dimensionality d_1 and d_2 is the rank of the matrices $\mathbf{G}$ and $\mathbf{W}_{\mathcal{G}}$ respectively. This establishes the relation between the triangle inequality of our metric and the positive semi–definitness of $\mathbf{G}$ and $\mathbf{W}_{\mathcal{G}}$ (please refer to Section 1).

5 Experimental results

In the experimental results, we validate the proposed metric in two different learning settings. First, we consider a supervised learning problem in the context of learning a linear transformation for dimensionality reduction. Second, we investigate the problem of unsupervised learning via spectral clustering algorithms where each entry in the affinity matrix uses the proposed kernel to measure the similarity between samples.

5.1 Supervised discriminative dimensionality reduction

Fisher/Linear discriminant analysis (FDA/LDA) seeks a low dimensional subspace where d_{KL} in Equation (2) is maximized [16]. For a 2–class/multi–class problem, FDA/LDA model each class as a Gaussian distribution under the assumption that all classes have equal covariance matrices. In this case, d_{KL} reduces to the Mahalanobis distance and FDA/LDA reduces to a GEP. To extend this framework when the covariance assumption does not hold, De La Torre and Kanade [7] proposed MODA that searches for a low dimen-

Table 1 Specifications of the data sets used in the discriminant analysis experiments where number of classes, size and the number of attributes are denoted by c, n and d respectively.

Data set	c	n	d	Data set	c	n	d
UCI Bupa	2	345	6	UCI Monks–III	2	554	6
UCI HouseVotes	2	435	16	UCI Pima	2	768	8
UCI Monks–I	2	556	6	UCI TicTacToe	2	958	9
UCI Monks–II	2	601	6				

sional subspace that explicitly maximizes the objective function in Equation (2). Our objective here is to use our separation measure in the same context and compare it to other discriminant analysis techniques.

Here, we only consider 2–class problems and model each class, C_1 and C_2 as a Gaussian distribution; $\mathcal{N}_1(\ \cdot\ ; \boldsymbol{\mu_1}, \boldsymbol{\Sigma_1})$ and $\mathcal{N}_2(\ \cdot\ ; \boldsymbol{\mu_2}, \boldsymbol{\Sigma_2})$ respectively. Similar to FDA/LDA and MODA, we search for a linear transformation $\mathbf{B} \in \mathbb{R}^{d \times k}$ with $k < d$ such that it maximizes $d_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2)$ in the lower dimensional space. The linear transformation $\mathbf{B}$ can have any number of bases k such that $1 \leq k \leq \min(d-1, n-1)$. This is unlike FDA/LDA which can only define subspaces of dimensionality $k \leq \min(c-1, d-1)$, where c is the number of classes.

Let the distance between $\mathcal{N}_1$ and $\mathcal{N}_2$ under the linear transformation $\mathbf{B}$ be defined as follows:

$$d_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2; \mathbf{B}) = \underbrace{\mathrm{tr}\{(\mathbf{B}^\top \mathbf{S} \mathbf{B})^{-1}(\mathbf{B}^\top \mathbf{U} \mathbf{B})\}}_{\mathrm{I}(\mathbf{B})} + \underbrace{\mathrm{tr}\{\log^2\{(\mathbf{B}^\top \boldsymbol{\Sigma_1} \mathbf{B})^{\mathbf{-1}}(\mathbf{B}^\top \boldsymbol{\Sigma_2} \mathbf{B})\}\}}_{\mathrm{II}(\mathbf{B})}, \tag{15}$$

where $\mathbf{U} = \mathbf{u}\mathbf{u}^\top$ and $\mathbf{S} = (\frac{1}{2}\boldsymbol{\Sigma_1} + \boldsymbol{\frac{1}{2}}\boldsymbol{\Sigma_2})$. Maximizing Equation (15) with respect to $\mathbf{B}$ yields a basis $\mathbf{B}^*_{\mathcal{G}}$ that is optimal, in terms of separation, for classes C_1 and C_2. Since there is no closed form solution for the maximum of Equation (15), we use an iterative procedure based on gradient ascent:

$$\mathbf{B}^{t+1} = \mathbf{B}^t + \eta \frac{\partial d_{\mathcal{G}}(\mathcal{N}_1, \mathcal{N}_2; \mathbf{B})}{\partial \mathbf{B}} = \mathbf{B}^t + \eta \frac{\partial \mathrm{I}(\mathbf{B})}{\partial \mathbf{B}} + \eta \frac{\partial \mathrm{II}(\mathbf{B})}{\partial \mathbf{B}},$$

where

$$\frac{\partial \mathrm{I}(\mathbf{B})}{\partial \mathbf{B}} = 2\mathbf{U}\mathbf{B}\boldsymbol{\Theta} - \mathbf{2}\mathbf{S}\mathbf{B}\boldsymbol{\Theta}(\mathbf{B}^\top \mathbf{U}\mathbf{B})\boldsymbol{\Theta}, \quad \boldsymbol{\Theta} = (\mathbf{B}^\top \mathbf{S}\mathbf{B})^{\mathbf{-1}},$$
$$\frac{\partial \mathrm{II}(\mathbf{B})}{\partial \mathbf{B}} = [2 \log \mathbf{L}\{2(\mathbf{B}^\top \boldsymbol{\Sigma_2}\mathbf{B})^{\mathbf{-1}}\mathbf{B}^\top \boldsymbol{\Sigma_2} - \mathbf{2}(\mathbf{B}^\top \boldsymbol{\Sigma_1}\mathbf{B})^{\mathbf{-1}}\mathbf{B}^\top \boldsymbol{\Sigma_1}\}]^\top,$$

with $\mathbf{L} = \mathrm{diag}(\ell_1, \ldots, \ell_k)$ is the eigenvalue matrix of $(\mathbf{B}^\top \boldsymbol{\Sigma_1}\mathbf{B})^{\mathbf{-1}}(\mathbf{B}^\top \boldsymbol{\Sigma_2}\mathbf{B})$, and η is the step length. The gradient ascent procedure starts with a reasonable step length and it is decreased by 50% if it increases the value of the

Table 2 Empirical error (with standard deviation) for discriminant analysis experiments using a projection dimension $k = 1$. Due to space limitations, please see supplementary material for $k = 2, 3$.

Data set	LDA	PCA+LDA	PCA	RCA	MODA	$\mathbf{B}^*_{\mathcal{G}}$
Bupa	44.7 (5.1)	37.3 (4.8)	45.5 (6.0)	37.9 (34.1)	34.1 (8.1)	**32.0** (6.2)
HouseVotes	11.1 (5.5)	4.5 (3.4)	12.6 (5.8)	4.5 (3.4)	4.5 (3.4)	**4.2** (3.6)
Monks–I	33.3 (8.6)	36.1 (10.5)	33.3 (8.6)	36.1 (10.5)	34.4 (10.5)	**33.1** (12.2)
Monks–II	37.9 (4.8)	33.7 (4.2)	43.9 (5.4)	35.0 (4.9)	32.5 (4.5)	**31.3** (4.7)
Monks–III	**18.8** (10.2)	22.4 (6.7)	34.2 (9.1)	22.4 (6.7)	23.3 (8.4)	21.1 (8.0)
pima	37.2 (5.1)	**24.0** (4.6)	39.6 (5.3)	**24.0** (4.8)	27.8 (4.5)	28.5 (5.9)
TicTacToe	38.9 (11.0)	**1.4** (4.6)	54.9 (9.6)	**1.4** (4.6)	**1.4** (4.5)	**1.5** (4.8)

objective function. Other strategies such as line search are possible but this simple method has provided good preliminary results. Similar to MODA, the objective function in Equation (15) is non–convex and any gradient ascent procedure can be trapped into local minima. Therefore, we typically start the algorithm with multiple initializations and select the solution $\mathbf{B}^*_{\mathcal{G}}$ with the lowest training error.

The error considered here is the error of a quadratic classifier in the lower dimensional space. Since each class is modelled as a Gaussian distribution, a sample $\mathbf{x}$ with an unknown label y is assigned the label of its closest class, where closeness is based on the Mahalanobis distance between the sample $\mathbf{x}$ and the class C_j: $(\mu_{\mathbf{j}} - \mathbf{x})^\top \boldsymbol{\Sigma}_{\mathbf{j}}^{\mathbf{-1}} (\mu_{\mathbf{j}} - \mathbf{x})$, $j = 1, 2$.

Table (1) shows seven data sets from the UCI ML Repository that are used in this experiment. The empirical error (with standard deviation) was averaged over 10 folds nested cross validation for three different projection dimensions $k = 1, 2, 3$. Table (2) shows the empirical error for LDA, PCA+LDA, PCA, RCA [2], MODA and $\mathbf{B}^*_{\mathcal{G}}$ on the UCI data sets for projection dimension $k = 1$. It is clear that linear transformation $\mathbf{B}^*_{\mathcal{G}}$ yields very competitive results with standard discriminant analysis techniques and with more recent approaches such as RCA and MODA.

5.2 Unsupervised clustering of images

In the second experiment, we consider an unsupervised learning problem where our main objective is to compare different distance and divergence measures between Gaussian densities in the context of clustering. Our hypothesis is that full metric measures such as $d_{\mathcal{G}}$ and d_H will yield better clustering results than d_{KL}, and very comparable to ρ.

Here, we adopt the same conceptual framework of Kondor and Jebara [15] that models each image as a bag of pixels (BOP). In this framework, instead of directly modelling each BOP as a Gaussian distribution, Kondor

Table 3 *Col. 1* The three data sets used in the spectral clustering experiments. *Col. 2* Number of classes, size and attributes for each data set. *Col. 3* The accuracy of spectral clustering using for the four similarity matrices.

Data set	c	n	d	$\mathbf{W}_B$	$\mathbf{W}_{KL}$	$\mathbf{W}_H$	$\mathbf{W}_{\mathcal{G}}$
Yale–A face data set	15	165	32×32	51.0	57.7	**64.9**	59.5
KTH TIPS grey scale textures	10	810	200×200	56.9	56.1	**60.5**	**60.2**
USPS handwritten digits	10	7291/2007	16×16	56.2	57.0	55.2	**59.1**

and Jebara map each BOP to a high dimensional feature space $\mathcal{H}$ using kernel PCA (KPCA) [23] in order to capture more nonlinear relations between the pixels. For KPCA, they use a Gaussian kernel with kernel width r. Next, they model each BOP in $\mathcal{H}$ as a Gaussian distribution with a full covariance and regularization parameter ϵ. That is, each image is finally represented as a Gaussian distribution in $\mathcal{H}$. Finally, they use SVMs with the Bhattacharyya kernel to classify the images. Please refer to [15] for more details. Here, we apply spectral clustering (SC) [18] on the Gaussian distributions in $\mathcal{H}$ instead of using SVMs.

Four similarity measures are used to construct the similarity or (adjacency) matrix for SC : $\mathbf{W}_B = \rho(\mathcal{N}_i, \mathcal{N}_j)$ – the Bhatacharyya kernel of [15], $\mathbf{W}_{KL} = \exp\{-d_{KL}(\mathcal{N}_i, \mathcal{N}_j)/\sigma\}$ – the KL kernel of [20], $\mathbf{W}_H = \exp\{-d_H(\mathcal{N}_i, \mathcal{N}_j)/\sigma\}$, and $\mathbf{W}_{\mathcal{G}} = \exp\{-d_{\mathcal{G}}(\mathcal{N}_i, \mathcal{N}_j)/\sigma\}$, where $\sigma > 0$ and $1 \leq i, j \leq n$. Note that, in total, there are three parameters to optimize in this framework; r, ϵ and σ. According to the recommendation in [18, pp. 6], we use the normalized SC of Shi and Malik [24]: $\mathbf{L}_{rw} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{W}$, where $\mathbf{W}$ can be replaced by any of the previously defined similarity matrices, $\mathbf{L}_{rw}$ is the normalized (as a random walk) Laplacian, and $\mathbf{D} = \text{diag}(\mathbf{W}\mathbf{1}_{n\times n})$. To proceed with SC, we find the c eigenvectors corresponding to the c smallest eigenvalues of the GEP: $\mathbf{L}_{rw}\mathbf{v} = \omega\mathbf{D}\mathbf{v}$ and form the matrix $\mathbf{V} = [\mathbf{v}_1 \ldots \mathbf{v}_c] \in \mathbb{R}^{n\times c}$, where c is the number of clusters. Now each Gaussian distribution in $\mathcal{H}$ (corresponding to one image) is mapped to a row vector in $\mathbf{V}$. Finally, we cluster the rows of $\mathbf{V}$ using the k–means algorithm – with multiple initializations – and select the clustering configuration with minimum distortion.

Three image data sets are used in these experiments and shown in Table (3). Due to the large size of the USPS data set, the first 100 digits of each class are considered as our data set. The number of clusters is assumed to be known and its equal to the number of classes in each data set. To measure the clustering accuracy, we adopt the technique in [28] that uses a $c \times c$ confusion matrix $\mathbf{C}$ and the Hungarian algorithm [14] to solve the following optimization problem: $\max \text{tr}\{\mathbf{CP}\}$, where $\mathbf{P}$ is a permutation matrix, and the result is divided by the number of data points to be clustered.

Table (3) shows the results of SC using the four different similarity measures. Due to the difficulty of clustering these images with a general and simple representation such as BOP, and due to sensitivity of this framework to the choice of parameter values as acknowledged by Kondor and Jebara, the accuracy is generally low for all the data sets. Nevertheless, we note the difference between ρ and d_{KL} on one hand, and d_H and $d_{\mathcal{G}}$ on the other hand. Counter to our hypothesis, the Bhattacharyya affinity ρ did not perform as good as the similarities induced by $d_{\mathcal{G}}$ and d_H.

The triangle inequality plays an important role for $\mathbf{W}$ and consequently for clustering. In the GEP of SC, $\mathbf{L}_{rw}$ should be PSD, $\mathbf{D}$ should be positive definite, and hence $\mathbf{W}$ should be PSD as well. If the triangle inequality is not satisfied, $\mathbf{W}$ will be non–definite and the GEP will yield an inaccurate embedding. It follows that the row vectors of $\mathbf{V}$ will have inaccurate coordinates, and consequently k–means will yield an inaccurate clustering. The amount of inaccuracy is tightly related to how far is $\mathbf{W}$ from a PSD matrix. This is where parameter σ comes into play for $\exp\{-d_{KL}/\sigma\}$ for instance, where it helped improve the positive semi–definiteness of $\mathbf{W}_{KL}$ thereby improving the clustering accuracy. A deeper and a more formal investigation is currently undergoing in this direction.

Concluding remarks : We have designed a metric that measures the separation or difference between two Gaussian densities. The measure has interesting properties and consequences for various learning algorithms and showed promising preliminary results in two different learning settings. Also, we have considered the importance of the triangle inequality axiom for metrics and divergence measures, and its relation to the PSD property of the gram matrix derived from these measures. Although our metric is a designed measure, an important and legitimate question to ask is, what is the original divergence measure between P_1 and P_2 such that when plugging in $\mathcal{N}_1$ and $\mathcal{N}_2$ yields our metric $d_{\mathcal{G}}$? The right answer is to generalize the analysis presented here using various divergence measures from the class of Aly–Silvey distances to the general form of the exponential family of probability distributions. On the one hand, it allows us to study which divergence measures factorize the difference in the exponential family in terms of difference in their statistics, and on the other hand, study which of these divergence measures satisfy the three metric axioms or yield symmetric PSD gram matrices. This analysis can result in a very rich set of measures that have different properties and characteristics, however this remains to be explored.

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. *Seris B* **28**(1), 131–142 (1966)
2. Bar-Hillel, A., Hertz, T., Shental, N., Weinshall, D.: Learning a Mahalanobis metric from equivalence constraints. J. of Machine Learning Research **6**, 937–965 (2005)
3. Bhattacharyya, A.: On a measure of divergence between two statistical populations defined by their probability distributions. Bull. Calcutta Math. Soc. **35**, 99–109 (1943)
4. Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics **22**, 493–507 (1952)

5. Christakos, G., Papanicolaou, V.: Norm–dependent covariance permissibility of weakly homogeneous spatial random fields and its consequences in spatial statistics. Stochastic Environmental Research and Risk Management **14**, 471–478 (2000)
6. Csiszár, I.: Information–type measures of difference of probability distributions and indirect observations. Studia Scientiarium Mathematicarum Hungarica **2**, 299–318 (1967)
7. De La Torre, F., Kanade, T.: Multimodal oriented discriminant analysis. In: ACM Proc. of ICML, pp. 177–184 (2005)
8. Elkan, C.: Using the trinagle inequality to accelerate k–means. In: ACM Proc. of ICML, pp. 147–153 (2003)
9. Förstner, W., Moonen, B.: A metric for covariance matrices. Tech. rep., Dept. of Geodesy and Geo–Informatics, Stuttgart University (1999)
10. Genton, M.: Classes of kernels for machine learning: A statistics perspective. J. of Machine Learning Research **2**, 299–312 (2001)
11. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: NIPS 11, pp. 487–493. MIT Press (1999)
12. Jebara, T., Kondor, R., Howard, A.: Probability product kenrels. J. of Machine Learning Research **5**, 819–844 (2004)
13. Kailath, T.: The divergence and Bhattacharyya distance measures in signal selection. IEEE Trans. on Communication Technology **15**(1), 52–60 (1967)
14. Knuth, D.E. (ed.): The Stanford graphbase. New York, Springer Verlag (1988)
15. Kondor, R., Jebara, T.: A kernel between sets of vectors. In: ACM Proc. of ICML (2003)
16. Kullback, S.: Information Theory and Statistics – Dover Edition. Dover, New York (1997)
17. Lafferty, J., Lebanon, G.: Information diffusion kernels. In: NIPS 14. MIT Press (2002)
18. Luxburg, U.v.: A tutotrial on spectral clustering. Tech. Rep. TR–149, Max Plank Institute for Biological Cybernetics (2006)
19. Martins, A., Smith, N., Xing, E., Aguiar, P., Figueiredo, M.: Nonextensive information theoretic kernels on measures. J. of Machine Learning Research **10**, 935–975 (2009)
20. Moreno, P., Ho, P., Vasconcelos, N.: A Kullback–Leibler divergence based kernel for svm classification in multimedia applications. In: NIPS 16 (2003)
21. Rao, C.: Use of Hellinger distance in graphical displays. In: E. Titt, T. Kollo, H. Niemi (eds.) Multivariate Statistics and Matrices in Statistics, pp. 143–161 (1995)
22. Roth, V., Laub, J., Buhmann, J.: Optimal cluster preserving embedding of nonmetric proximity data. IEEE Trans. PAMI **25**(12), 1540–1551 (2003)
23. Schölkopf, B., Smola, A.: Learning with kernels. MIT Press, Cambridge, MA (2002)
24. Shi, J., Malik, J.: Motion segmentation and tracking using normalized cuts. In: IEEE Proc. of ICCV, pp. 1154–1160 (1998)
25. Sriperumbudur, B., Gretton, A., Fukumizu, K., Schölkopf, B., Lanckriet, G.: Hilbert space embeddings and metrics on probability distributions. J. of Machine Learning Research **11**, 1517–1561 (2010)
26. Tsuda, K., Kawanabi, M., Ratsch, G., Sonnenburg, S., Muller, K.R.: A new discriminative kernel from probability distributions. Neural Computation **14**, 2397–2414 (2002)
27. Young, G., Householder, A.: Discussion of a set of points in temrs of their mutual distances. Psychometrika **3**(1), 19–22 (1938)
28. Zha, H., Ding, C., Gu, M., He, X., Simon, H.: Spectral relaxation for k–means clustering. In: NIPS 13. MIT Press (2001)