2013 IEEE 13th International Conference on Data Mining

# Efficient and Scalable Information Geometry Metric Learning

Wei Wang*†, Bao-Gang Hu†, Zengfu Wang*
*Department of Automation, University of Science and Technology of China, Hefei, China
†Institute of Automation, Chinese Academy of Sciences, Beijing, China
*†weiyi889@mail.ustc.edu.cn, †hubg@nlpr.ia.ac.cn, *zfwang@ustc.edu.cn

***Abstract*—Information Geometry Metric Learning (IGML) is shown to be an effective algorithm for distance metric learning. In this paper, we attempt to alleviate two limitations of IGML: (A) the time complexity of IGML increases rapidly for high-dimensional data; (B) IGML has to transform the input low-rank kernel into a full-rank one since it is undefined for singular matrices. To this end, two novel algorithms, referred to as Efficient Information Geometry Metric Learning (EIGML) and Scalable Information Geometry Metric Learning (SIGML), are proposed. EIGML scales linearly with the dimensionality, resulting in significantly reduced computational complexity. As for SIGML, it is proven to have a *range-space preserving* property. Following this property, SIGML is found to be capable of handling both full-rank and low-rank kernels. Additionally, the geometric information from data is further exploited in SIGML. In contrast to most existing metric learning methods, both EIGML and SIGML have closed-form solutions and can be efficiently optimized. Experimental results on various data sets demonstrate that the proposed methods outperform the state-of-the-art metric learning algorithms.**

***Keywords*—Mahalanobis distance learning, Information Geometry Metric Learning, closed-form solution, range-space preserving**

## I. Introduction

Metric learning is of fundamental interest in machine learning and data mining. The metric distances provide a measurement of dissimilarity between different points and have a critical impact on the success of many classical algorithms, e.g., $k$-nearest neighbor (kNN) classification, support vector machines (SVM) and radial basis function (RBF) networks. To this end, a number of excellent methods have been developed for exploiting distance metrics in different settings, e.g., [1], [2], [3], [6]. In this direction, much representative work tends to learn a global Mahalanobis distance from labeled data [3], [4], [5], [11], [13]. These global methods formulate a semidefinite programming (SDP) [24] to keep all the data points in the same class close together while ensuring those from different classes far apart. To further utilize both the label information and the geometric information from data, some local Mahalanobis distance learning algorithms are proposed recently [7], [8], [9], [10] which can be divided into two categories: convex and nonconvex. The convex method [8] optimizes an SDP and perform a full eigen-decomposition of the Mahalanobis distance at each iteration to maintain the positive semi-definite property. Therefore, the computational cost rises rapidly with the increase of dimension. The nonconvex methods [7], [9], [10] are prone to being trapped in local solutions and may suffer from computationally expensive optimizations.

In contrast to the metric learning methods discussed above, Information Geometry Metric Learning (IGML), introduced by Wang & Jin [12], can find a closed-form solution rather than solving an SDP. Despite the popularity of IGML in metric learning, it may show some limitations in the following aspects. (1) The computational complexity of IGML is $O(d^3 + nd^2)$, where $n$ is the number of training points and $d$ is the dimensionality. The running time increases rapidly for high-dimensional data. (2) IGML employs a (potential) low-rank ideal kernel [13]. However, IGML requires that the input kernel should be full-rank, since it is infinite for singular matrices. Restricted by this requirement, IGML has to smooth the low-rank kernel with an identity matrix and a smoother parameter. This smoother has great influences on metric learning performance. Moreover, the ideal kernel used in IGML only exploits the label information, thus the geometric information of data is lost.

In this paper, we attempt to alleviate the limitations discussed above respectively and propose two novel distance metric learning algorithms, i.e., Efficient Information Geometry Metric Learning (EIGML) and Scalable Information Geometry Metric Learning (SIGML). EIGML is advantageous for high-dimensional data sets since it reduces the computational complexity of IGML to $O(nd)$ and is also fast for testing. In particular, EIGML learns an identity plus low-rank Mahalanobis distance [16] and represents the Mahalanobis distance by $O(d)$ parameters instead of $O(d^2)$ memory. On the other hand, SIGML is proposed to generalize and formalize IGML to both full-rank and low-rank kernel cases. We prove that SIGML has a *range-space preserving* property. That is, SIGML is finite if the range space of the kernel matrix contains the range space of the Mahalanobis distance metric and the Mahalanobis distance metric is positive definite. Based on this property, for low-rank kernels, SIGML is restricted to the range spaces of the matrices and alternatively solves a full-rank problem in a lower dimensional space. In this term, SIGML can directly handle the low-rank kernels without smoother parameter. Another key contribution of SIGML is that the learnt Mahalanobis distance simultaneously integrates two sources of information: 1) preserving the neighborhood structure according to data itself; 2) maximally aligning with the labels of data. We emphasize that both EIGML and SIGML can find the closed-form solutions, leading to efficient optimization.

The rest of paper is organized as follows. IGML is reviewed briefly in Section II. Section III presents EIGML for high dimensional data. In Section IV, SIGML is proposed to expand the domain of IGML. We also show how to effectively preserve

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.67

the local information in SIGML. Section V reports the experimental results. Finally, conclusions are given in Section VI.

## II. Relative Formulas

In this section, we outline Information Geometry Metric Learning (IGML) [12]. Associated with two covariance matrices $\mathbf{P}$ and $\mathbf{Q}$, two Gaussian distributions with zero mean are defined as: $Pr(\mathbf{x}|\mathbf{P})$ and $Pr(\mathbf{x}|\mathbf{Q})$.

**Theorem 1.** [12] *The distance function between two positive definite matrices $\mathbf{P}$ and $\mathbf{Q}$ can be derived by the Kullback-Leibler divergence between $Pr(\mathbf{x}|\mathbf{P})$ and $Pr(\mathbf{x}|\mathbf{Q})$ which is equal to the following Bregman distance function [15]:*

$$d(\mathbf{P}\|\mathbf{Q}) = \frac{1}{2}(tr(\mathbf{Q}^{-1}\mathbf{P}) + log|\mathbf{Q}| - log|\mathbf{P}| - n). \quad (1)$$

Given a set of variables $\mathbf{X} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_2)\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i$ is the class label, the Mahalanobis distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ can be calculated as follows:

$$d_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j), \quad (2)$$

where $\mathbf{A} \in \mathbb{R}^{d\times d}$ is positively semi-definite. IGML constructs a linear kernel $\mathbf{K_X}$ for $\mathbf{A}$: $\mathbf{K_X} = \mathbf{X}^T\mathbf{A}\mathbf{X}$. In the ideal kernel function, two points should be considered similar if and only if they belong to the same class. Thus, the so-called *ideal kernel* [13] is given by:

$$\mathbf{K}_D(i,j) = \mathbf{K}_D(\mathbf{x}_i, \mathbf{x}_j) = \begin{cases} 1 & \text{if } y_i = y_j \\ 0 & \text{if } y_i \neq y_j. \end{cases} \quad (3)$$

When the number of classes $C < n$, $\mathbf{K}_D$ is singular. Therefore, $\mathbf{K}_D$ has to be smoothed with an identify matrix $\mathbf{I}_n$, as: $\bar{\mathbf{K}}_D = \mathbf{K}_D + \lambda\mathbf{I}_n$, where $\lambda > 0$ is the smoother parameter. In IGML, the optimal distance metric $\mathbf{A}$ is searched by minimizing the matrix distance $d(\mathbf{K_X}\|\bar{\mathbf{K}}_\mathbf{D})$ defined in Eq. (1):

$$\begin{aligned} \mathbf{A} &= arg\min_{\mathbf{A}\succeq 0} d(\mathbf{K_X}\|\bar{\mathbf{K}}_\mathbf{D}) \\ &= arg\min_{\mathbf{A}\succeq 0} tr(\bar{\mathbf{K}}_D^{-1}\mathbf{X}^T\mathbf{A}\mathbf{X}) - log|\mathbf{A}|. \end{aligned} \quad (4)$$

**Proposition 1.** [12] *The optimal solution to Eq. (4) is*

$$\mathbf{A} = (\mathbf{X}\bar{\mathbf{K}}_D^{-1}\mathbf{X}^T)^{-1}. \quad (5)$$

*The optimal $\mathbf{A}$ in Eq. (5) can also be expressed as follows:*

$$\mathbf{A} = \lambda(\sum_{k=1}^{C} s_k[\mathbf{\Sigma}_k + \frac{\lambda\bar{\mathbf{x}}_k\bar{\mathbf{x}}_k^T}{\lambda + s_k}])^{-1}, \quad (6)$$

*where $s_k$ is the number of the input points in the $k$-th class, $\bar{x}_k$ and $\Sigma_k$ are the mean and the covariance matrix for the points in the $k$-th class respectively.*

## III. Efficient Information Geometry Metric Learning

Note that the computational complexity of IGML is $O(d^3 + nd^2)$, relying on the computation of covariance matrix and matrix inversion in Eq. (6). Moreover, the number of involved parameters is $O(d^2)$. For example, a data set with 1,000 dimensions leads to a Mahalanobis distance matrix with 1 million parameters. These aspects limit the application of IGML in high dimensional data. With the purpose of improving computational efficiency, we propose Efficient Information Geometry Metric Learning (EIGML).

Consider a low-dimensional space in $\mathbb{R}^d$ and let the columns of $\mathbf{U}$ form an orthogonal basis of this subspace. The Mahalanobis distance $\mathbf{A}$ is constrained to take the form [16]: $\mathbf{A} = \mathbf{I}^d + \mathbf{A}_l = \mathbf{I}^d + \mathbf{U}\mathbf{L}\mathbf{U}^T$, where $\mathbf{I}^d \in R^{d\times d}$ is the identity matrix, $\mathbf{A}_l$ denotes the low-rank part of $\mathbf{A}$ and $\mathbf{L} \in \mathbb{S}_+^{k\times k}$ ($k \ll d$). The selection of $\mathbf{U}$ can follow some heuristics in [16], [17]. We propose the following algorithm EIGML to learn an identity plus low-rank Mahalanobis distance:

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{L}\succeq 0} tr(\bar{\mathbf{K}}_D^{-1}\mathbf{X}^T\mathbf{A}\mathbf{X}) - log|\mathbf{A}| \\ &s.t.\ \ \mathbf{A} = \mathbf{I}^d + \mathbf{U}\mathbf{L}\mathbf{U}^T. \end{aligned} \quad (7)$$

Following the Sylvester's determinant lemma [18], the second term of Eq. (7) can be rewritten as:

$$log|\mathbf{A}| = log|\mathbf{I}^d + \mathbf{U}\mathbf{L}\mathbf{U}^T| = log|\mathbf{I}^k + \mathbf{L}| = log|\mathbf{S}|, \quad (8)$$

where $\mathbf{S} = \mathbf{I}^k + \mathbf{L}$. The first term of Eq. (7) is rewritten as:

$$\begin{aligned} &tr(\bar{\mathbf{K}}_D^{-1}\mathbf{X}^T\mathbf{A}\mathbf{X}) = tr((\mathbf{I}^d + \mathbf{U}\mathbf{L}\mathbf{U}^T)\mathbf{X}\bar{\mathbf{K}}_D^{-1}\mathbf{X}^T) \\ &= tr(\mathbf{X}\bar{\mathbf{K}}_D^{-1}\mathbf{X}^T) + tr(\mathbf{L}\mathbf{U}^T\mathbf{X}\bar{\mathbf{K}}_D^{-1}\mathbf{X}^T\mathbf{U}) \\ &= tr(\mathbf{X}\bar{\mathbf{K}}_D^{-1}\mathbf{X}^T) - tr(\mathbf{X}'\bar{\mathbf{K}}_D^{-1}\mathbf{X}'^T) + tr(\mathbf{S}\mathbf{X}'\bar{\mathbf{K}}_D^{-1}\mathbf{X}'^T), \end{aligned} \quad (9)$$

where $\mathbf{X}' = \mathbf{U}^T\mathbf{X}$ can be considered as the reduced-dimensional representation of $\mathbf{X}$.

Substituting Eq. (8) and Eq. (9) into Eq. (7), we obtain the following objective function:

$$\min_{\mathbf{S}\succeq 0} tr(\bar{\mathbf{K}}_D^{-1}\mathbf{X}'^T\mathbf{S}\mathbf{X}') - log|\mathbf{S}|. \quad (10)$$

The solution of Eq. (10) can be given as follows based on Proposition 1:

$$\mathbf{S} = \lambda(\sum_{k=1}^{C} s_k[\mathbf{\Sigma}'_k + \frac{\lambda\bar{\mathbf{x}}'_k\bar{\mathbf{x}}'^T_k}{\lambda + s_k}])^{-1}, \quad (11)$$

where $\bar{x}'_k$ and $\Sigma'_k$ are the mean and the covariance matrix for the points of $\mathbf{X}'$ in the $k$-th class respectively. Note that Eq. (11) solves for a $k \times k$ matrix rather than a $d \times d$ matrix required in IGML. The computational complexity is reduced to $O(k^3 + nk^2 + ndk)$. Based on Eq. (7), in EIGML, the optimal Mahalanobis distance $\mathbf{A}$ is obtained as: $\mathbf{A} = \mathbf{I}^d + \mathbf{U}(\mathbf{S} - \mathbf{I}^k)\mathbf{U}^T$. EIGML also shows its advantages in less testing time since it can store the optimal $\mathbf{A}$ implicitly using $O(dk + k^2)$ memory and the Mahalanobis distance between any two points can be computed in $O(dk + k^2)$ time.

## IV. Scalable Information Geometry Metric Learning

Recall that the ideal kernel $\mathbf{K}_D$ defined in Eq. (3) is (potential) low-rank. IGML modifies $\mathbf{K}_D$ to the full-rank $\bar{\mathbf{K}}_D$ using an identity matrix since IGML is undefined and infinite for low-rank kernel matrices. However, this modification may change some intrinsic properties of $\mathbf{K}_D$ and the involved smoother parameter has great influences on metric learning performance. Moreover, $\mathbf{K}_D$ is derived only from the labels of data while neglects the local structure information. Generally, we extend the idea of IGML and propose the following

optimization problem, named Scalable Information Geometry Metric Learning (SIGML):

$$\mathbf{A} = arg\min_{\mathbf{A}\geq 0} tr(\mathbf{K}_I^{-1}\mathbf{X}^T\mathbf{A}\mathbf{X}) - log|\mathbf{A}|, \quad (12)$$

where $\mathbf{K}_I$ is the ideal kernel constructed by exploiting more information from data. Meanwhile, $\mathbf{K}_I$ can be either full-rank or low-rank, i.e., rank$(\mathbf{K}_I) = r \leq n$.

### A. Range-Space Preserving Property

We relate the optimization defined in Eq. (12) to the real-valued strictly convex function $\phi$ over a convex set.

**Proposition 2.** *The objective function in Eq. (12) is equivalent to the following function:*

$$tr(\mathbf{K}_I^{-1}\mathbf{X}^T\mathbf{A}\mathbf{X}) - log|\mathbf{A}| = \phi(\mathbf{A}) - tr(\mathbf{A}-\mathbf{K}_I)^T\nabla\phi(\mathbf{K}_I), \quad (13)$$

*where* $\phi(\mathbf{X}) = -log|\mathbf{X}| = -\sum_i log\lambda_i = \sum_i \phi(\lambda_i)$.

Based on the eigenvalues and eigenvectors of $\mathbf{A}$ and $\mathbf{K}_I$, an alternative expression for Eq. (13) can be provided in the following theorem.

**Theorem 2.** *Let the eigen-decomposition of* $\mathbf{A}$ *and* $\mathbf{K}_I$ *be* $\mathbf{V}\Lambda\mathbf{V}^T$ *and* $\mathbf{U}\Theta\mathbf{U}^T$*, respectively. Then* $\mathbf{X}^T\mathbf{A}\mathbf{X} = \mathbf{X}^T\mathbf{V}\Lambda\mathbf{V}^T\mathbf{X} = \mathbf{M}\Lambda\mathbf{M}^T$*, where* $\mathbf{M} = \mathbf{X}^T\mathbf{V}$*. Eq. (13) is equal to the following expression:*

$$\sum_{i=1}^{d}\sum_{j=1}^{n}(\mathbf{m}_i^T\mathbf{u}_j)^2(-\frac{log\lambda_i}{|\mathbf{m}_i|^2} + \frac{\lambda_i}{\theta_j}) - n. \quad (14)$$

*Proof:*

$$\begin{aligned}
&F(\mathbf{A}) = \phi(\mathbf{A}) - tr(\mathbf{A}-\mathbf{K}_I)^T\nabla\phi(\mathbf{K}_I)\\
&= \sum_{i=1}^{d}\phi(\lambda_i) - tr((\mathbf{M}\Lambda\mathbf{M}^T - \mathbf{U}\Theta\mathbf{U}^T)^T\nabla\phi(\mathbf{K}_I))\\
&= \sum_{i=1}^{d}\sum_{j=1}^{n}(\mathbf{m}_i^T\mathbf{u}_j)^2\frac{\phi(\lambda_i)}{|\mathbf{m}_i|^2} - tr((\mathbf{M}\Lambda\mathbf{M}^T - \mathbf{U}\Theta\mathbf{U}^T)^T\nabla\phi(\mathbf{K}_I)),
\end{aligned}$$

where the third line uses the fact that $\sum_{j=1}^{n}(\mathbf{m}_i^T\mathbf{u}_j)^2 = |\mathbf{m}_i|^2$. $\nabla\phi(\mathbf{K}_I)$ can be rewritten as the following function [22]:

$$\nabla\phi(\mathbf{K}_I) = \mathbf{U}\begin{pmatrix}\nabla\phi(\theta_1) & 0 & \cdots\\ 0 & \nabla\phi(\theta_2) & \cdots\\ \vdots & \vdots & \ddots\end{pmatrix}\mathbf{U}^T.$$

In term of this notation, we have:

$$\begin{aligned}
&tr(\mathbf{U}\Theta\mathbf{U}^T\nabla\phi(\mathbf{K}_I)) = \sum_{j=1}^{n}\theta_j\nabla\phi(\theta_j),\\
&tr(\mathbf{M}\Lambda\mathbf{M}^T\nabla\phi(\mathbf{K}_I)) = \sum_{i=1}^{d}\sum_{j=1}^{n}(\mathbf{m}_i^T\mathbf{u}_j)^2\lambda_i\nabla\phi(\theta_j).
\end{aligned}$$

Finally, $F(\mathbf{A})$ is equal to the following expression:

$$\sum_{i=1}^{d}\sum_{j=1}^{n}(\mathbf{m}_i^T\mathbf{u}_j)^2(\frac{\phi(\lambda_i)}{|\mathbf{m}_i|^2} - \lambda_i\nabla\phi(\theta_j)) + \sum_{j=1}^{n}\theta_j\nabla\phi(\theta_j).$$

Substituting $\phi(x) = -log|x|$ into $F(\mathbf{A})$, we obtain:

$$F(\mathbf{A}) = \sum_{i=1}^{d}\sum_{j=1}^{n}(\mathbf{m}_i^T\mathbf{u}_j)^2(-\frac{log\lambda_i}{|\mathbf{m}_i|^2} + \frac{\lambda_i}{\theta_j}) - n.$$

■

In the information geometry metric learning problems, the optimal Mahalanobis distance $\mathbf{A}$ is expected to be full-rank. In Eq. (14), if $\mathbf{K}_I$ is of low rank, some eigenvalues $\theta_i$ are equal to zero. However, $F(\mathbf{A})$ is infinite when $\lambda_i \neq 0$ but $\theta_i = 0$. [22] proves that the key to using $\phi(\mathbf{X}) = -log|\mathbf{X}|$ in the low-rank setting comes from restricting $\phi(\mathbf{X})$ to the range spaces of matrices. To make $F(\mathbf{A})$ finite for rank-deficient $\mathbf{K}_I$, we have the following Proposition 3.

**Proposition 3.** *The objective function in Eq. (14) is finite if rank(*$\mathbf{A}$*)=*$d$ *and rank(*$\mathbf{A}$*)* $\leq$ *rank(*$\mathbf{K}_I$*).*

*Proof:* For finiteness of $F(\mathbf{A})$, we have that $\mathbf{m}_i$ and $\mathbf{u}_j$ must be orthogonal (i.e., $\mathbf{m}_i^T\mathbf{u}_j = 0$) when $\lambda_i \neq 0$ but $\theta_i = 0$. $\mathbf{m}_i$ is also the eigenvector of $\mathbf{A}$ since $\mathbf{m}_i = \mathbf{X}^T\mathbf{v}_i$. When $\theta_i = 0$, the corresponding eigenvector $\mathbf{u}_j$ is in the null space of $\mathbf{K}_I$. When $\lambda_i \neq 0$, the corresponding eigenvector $\mathbf{m}_i$ is in the range space of $\mathbf{A}$. Therefore, every vector $\mathbf{u}_j$ in the null space of $\mathbf{K}_I$ is orthogonal to any vector $\mathbf{m}_i$ in the range space of $\mathbf{A}$. This explains that Null$(\mathbf{K}_I) \subseteq$ Null$(\mathbf{A})$ or, equivalently Range$(\mathbf{A}) \subseteq$ Range$(\mathbf{K}_I)$. The property that Range$(\mathbf{A}) \subseteq$ Range$(\mathbf{K}_I)$ implies rank$(\mathbf{A}) \leq$ rank$(\mathbf{K}_I)$. ■

Based on Theorem 2 and Proposition 3, SIGML for low-rank matrices can be expressed as:

$$F(\mathbf{A}) = \sum_{i=1}^{d}\sum_{j=1}^{r}(\mathbf{m}_i^T\mathbf{u}_j)^2(-\frac{log\lambda_i}{|\mathbf{m}_i|^2} + \frac{\lambda_i}{\theta_j}) - n, \quad (15)$$

where $r =$ rank$(\mathbf{K}_I)$ and $d \leq r < n$. The eigenvalues of $\mathbf{A}$ and $\mathbf{K}_I$ are listed in non-increasing order.

### B. Low-Rank SIGML via Restriction on the Range Space

The section above demonstrates that the range space of $\mathbf{K}_I$ must contain the range space of $\mathbf{A}$ in the low-rank cases. We now show that the optimization problem (15) for low-rank $\mathbf{K}_I$ can be cast as a full rank problem in a lower dimensional space (i.e., the range space of $\mathbf{K}_I$). Afterwards, the closed-form solution to SIGML can be given.

**Theorem 3.** *Let the positive semidefinite* $d \times d$ *matric* $\mathbf{A}$ *and* $n \times n$ *matric* $\mathbf{K}_I$ *satisfy rank*$(\mathbf{A}) = d$ *and rank*$(\mathbf{K}_I) = r$*, Range(*$\mathbf{A}$*)* $\subseteq$ *Range(*$\mathbf{K}_I$*). Let* $\mathbf{W}$ *be an* $n \times r$ *column orthogonal matrix with Range(*$\mathbf{K}_I$*)* $\subseteq$ *Range(*$\mathbf{W}$*). If we define:*

$$\begin{aligned}
&tr(\mathbf{K}_I^{-1}\mathbf{X}^T\mathbf{A}\mathbf{X}) - log|\mathbf{A}|\\
&= tr((\mathbf{W}^T\mathbf{K}_I\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{W}) - log|\mathbf{A}|.
\end{aligned} \quad (16)$$

*Then this definition is consistent with the low-rank IGML in Eq. (15) and is not dependent on the choice of* $\mathbf{W}$*.*

*Proof:* Denote $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^T$ and $\mathbf{K}_I = \mathbf{U}\Theta\mathbf{U}^T$, respectively. Assume the eigenvalues of $\mathbf{A}$ and $\mathbf{K}_I$ are sorted in non-increasing order. The upper left $r \times r$ submatrix of $\Theta$

are $\Theta_r$, and the corresponding reduced eigenvectors are $\mathbf{U}_r$. Substituting $\mathbf{W} = \mathbf{U}_r$ into Eq. (16), we obtain:

$$\begin{aligned} &tr((\mathbf{U}_r^T\mathbf{U}\Theta\mathbf{U}^T U_r)^{-1}\mathbf{U}_r^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{U}_r) - log|\mathbf{A}| \\ &= tr(\Theta_r^{-1}\mathbf{U}_r^T\mathbf{X}^T\mathbf{V}\Lambda\mathbf{V}^T\mathbf{X}\mathbf{U}_r) - log|\mathbf{A}| \\ &= tr(\Theta_r^{-1}(\mathbf{U}_r^T\mathbf{M})\Lambda(\mathbf{M}^T\mathbf{U}_r)) - log|\mathbf{A}|, \end{aligned}$$

where $\Theta_r$ and $\mathbf{A}$ are full-rank. Therefore, following Theorem 2, Eq. (16) can be written as:

$$\sum_{i=1}^{d}\sum_{j=1}^{r}(\mathbf{m}_i^T\mathbf{u}_j)^2(-\frac{log\lambda_i}{|\mathbf{m}_i|^2} + \frac{\lambda_i}{\theta_j}) - n.$$

Next, we prove that this definition is independent of the choice of $\mathbf{W}$. Note that all the $n \times r$ orthogonal matrices with the same range space of $\mathbf{U}_r$ can be expressed as $\mathbf{WQ}$, where $\mathbf{Q}$ is an $r \times r$ orthogonal matrix. Substituting $\mathbf{WQ}$ into $\mathbf{W}$ in Eq. (16), we obtain:

$$\begin{aligned} &tr(((\mathbf{WQ})^T\mathbf{K}_I(\mathbf{WQ}))^{-1}(\mathbf{WQ})^T\mathbf{X}^T\mathbf{A}\mathbf{X}(\mathbf{WQ})) \\ &= tr(\mathbf{Q}^{-1}(\mathbf{W}^T\mathbf{K}_I\mathbf{W})^{-1}\mathbf{Q}^{-T}\mathbf{Q}^T(\mathbf{W}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{W})\mathbf{Q}) \\ &= tr((\mathbf{W}^T\mathbf{K}_I\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{W}), \end{aligned}$$

where the third line uses the fact $\mathbf{Q}$ and $\mathbf{W}^T\mathbf{K}_I\mathbf{W}$ are both square, non-singular matrices. ■

Following Theorem 3 and Proposition 1, the solution of SIGML in Eq. (12) can be given by:

$$\mathbf{A} = \begin{cases} (\mathbf{X}\mathbf{K}_I^{-1}\mathbf{X}^T)^{-1} & \text{if } r = n \\ (\mathbf{X}\mathbf{W}(\mathbf{W}^T\mathbf{K}_I\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{A})^{-1} & \text{if } d \le r < n, \end{cases} \quad (17)$$

where $\mathbf{W}$ is an $n \times r$ column orthogonal matrix with Range($\mathbf{K}_I$) $\subseteq$ Range($\mathbf{W}$).

### C. Locality Preserving in SIGML

In IGML, the optimal Mahalanobis metric is learnt based on $\mathbf{K}_D$ for classification. $\mathbf{K}_D$ is formulated using only the class labels of data points and has the fixed rank $C$. It is expected to construct an ideal kernel $\mathbf{K}_I$ which has an adjustable rank (full rank or low rank). Moreover, the construction of $\mathbf{K}_I$ should be based on two simple idealizations for KNN classification: 1) similarities between points with different labels will be zero; 2) similarities between points in the same class will be encouraged according to their relative locations. In this term, $\mathbf{K}_I$ in SIGML is defined on a weighted graph structure $\mathbf{G}$ with the following adjacency matrix $\mathbf{M}$:

$$\mathbf{M}(i,j) = \begin{cases} exp(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}) & \text{if } i \ne j \ \&\ y_i = y_j \\ 0 & otherwise, \end{cases} \quad (18)$$

where $\sigma$ is the width of Gaussian function.

*1) Full-Rank Ideal Kernel $\mathbf{K}_I$:* Specific kernel functions can be constructed to capture a useful and more global sense of similarity on the graph $\mathbf{G}$ [19], [20]. Let $\mathbf{D}$ be an $n \times n$ diagonal matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{M}_{ij}$. The Laplacian of $\mathbf{G}$ can be defined as $\mathbf{L} = \mathbf{D} - \mathbf{M}$, and the Normalized Laplacian is $\tilde{\mathbf{L}} = \mathbf{D}^{-\frac{1}{2}}\mathbf{L}\mathbf{D}^{-\frac{1}{2}}$. The eigenvalues and eigenvectors of $\tilde{\mathbf{L}}$ are denoted as $\lambda_i$ and $\phi_i$, so that $\tilde{\mathbf{L}} = \sum_i \lambda_i\phi_i\phi_i^T$. In this paper, we investigate the diffusion kernel [19] which is proven to be a generalization of Gaussian kernel to graphs. In this term, $\mathbf{K}_I$ is defined as: $\mathbf{K}_I = \sum_{i=1}^{n} exp(-\sigma_d^2/2\lambda_i)\phi_i\phi_i^T$, where $\sigma_d$ is the width of the diffusion kernel.

*2) Low-Rank Ideal Kernel $\mathbf{K}_{LI}$:* The kernel matrix $\mathbf{K}_I$ constructed above is of full rank since its eigenvalues $f(\lambda_i) = exp(-\sigma_d^2/2\lambda_i) > 0$ $(i = 1, ..., n)$. Despite the effectiveness of $n \times n$ matrix $\mathbf{K}_I$, it has relatively high computational efficiency due to the full eigen-decomposition of Laplacian $\tilde{\mathbf{L}}$ and the matrix multiplication. Recently, many kernel-based algorithms [21], [22] employ low-rank representations to improve computational efficiency. Recall that $\tilde{\mathbf{L}}$ has an interesting property [20]: a large $\lambda_i$ corresponds to a rather uneven $\phi_i$ on the graph $\mathbf{G}$. From a regularization perspective, the uneven vectors should be penalized more strongly than the vectors with small eigenvalues, since the changes should vary slowly over the graph [20]. Based on this intention, we can construct a low-rank $\mathbf{K}_{LI}$ (i.e., rank($\mathbf{K}_{LI}$) $= r \le n$) which just contains only the $r$ smallest eigenvalues $\lambda_i$ and the corresponding eigenvector $\phi_i$: $\mathbf{K}_{LI} = \sum_{i=1}^{r} exp(-\sigma_d^2/2\lambda_i)\phi_i\phi_i^T$, where the eigenvalues of $\tilde{\mathbf{L}}$ are listed in non-decreasing order. This low-rank representation relieves the burden of memory from $O(n^2)$ storage to $O(nr)$ and has the advantages of efficient eigen-decomposition and fast matrix multiplication.

### D. Proposed Algorithms

For full-rank $\mathbf{K}_I$ in SIGML (SIGML-F), we can find the following closed-form solution based on Eq. (17):

$$\mathbf{A} = (\mathbf{X}\mathbf{K}_I^{-1}\mathbf{X}^T)^{-1} = (\mathbf{X}\Phi\Lambda^{-1}\Phi^T\mathbf{X}^T)^{-1}, \quad (19)$$

where $\mathbf{\Phi} = (\phi_1, ..., \phi_n)$ and $\Lambda^{-1}$ is an $n \times n$ diagonal matrix with $[\Lambda^{-1}]_{ii} = [exp(-\sigma_d^2/2\lambda_i)]^{-1}$.

For low-rank $\mathbf{K}_{LI}$ in SIGML (SIGML-L), we can find the following closed-form solution based on Eq. (17):

$$\begin{aligned} \mathbf{A} &= (\mathbf{W}^T\mathbf{K}_{LI}\mathbf{W})^{-1}\mathbf{W}^T\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{W} \\ &= (\mathbf{X}\mathbf{W}\Lambda_r^{-1}\mathbf{W}^T\mathbf{X}^T)^{-1}, \end{aligned} \quad (20)$$

where $\mathbf{W}$ is an orthogonal $n \times r$ matrix: $\mathbf{W} = (\phi_1, \phi_2, ..., \phi_r)$ and $\Lambda_r^{-1}$ is a $r \times r$ diagonal matrix with $[\Lambda_r^{-1}]_{ii} = [exp(-\sigma_d^2/2\lambda_i)]^{-1}$.

## V. Experiments

In this section, we evaluate the proposed metric learning methods (i.e., EIGML and SIGML) for classification tasks. To quantitatively evaluate the separability of points in different classes, a simple measurement is used: misclassification rate by 1-nearest neighbor classifier (1-NN). The experiments are presented on 16 data sets from different domains: UCI data[1], IDA data[2], text data 20newsgroups[3] and handwritten digit data USPS[3]. Their detailed information can be found in Table I. Each data set is randomly split into training and test subsets and all metrics are trained using only the training sets. Test instances are compared to points in the training sets using the learnt distance metric. Results are obtained by averaging 100 runs on random splits of these data sets. For all the data sets, the features are scaled into $[-1, +1]$. The experiments are carried out on a single machine with Intel Core 2 Quad @ 2.40 Ghz and 10 GB of RAM running 64-bit Windows 7.

[1]Available at: urlhttp://www.ics.uci.edu/~mlearning/.

[2]Available at: http://ida.first.fraunhofer.de/projects/bench/benchmarks.html.

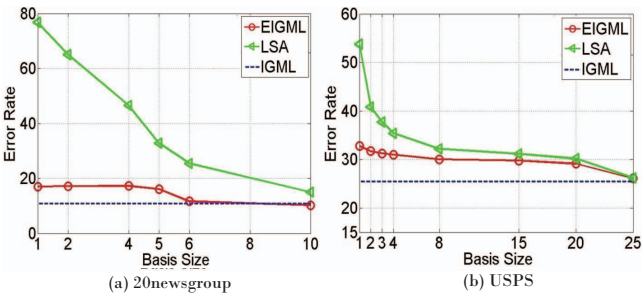
[3]Available at: http://www.cs.nyu.edu/~roweis/data.html.

Table I. LIST OF DATA SETS. DATA SETS WITH * CONTAIN INTRINSIC WITHIN-CLASS MULTIMODAL STRUCTURES.

| Data name | # of attributes | Training Size | Test Size | Classes |
|---|---|---|---|---|
| 20newsgroups | 100 | 11370 | 4872 | 4 |
| USPS | 784 | 7700 | 3300 | 10 |
| IDA: banana* | 2 | 400 | 4900 | 2 |
| IDA: breast-cancer | 9 | 200 | 77 | 2 |
| IDA: diabetes | 8 | 468 | 300 | 2 |
| IDA: flare-solar | 9 | 666 | 400 | 2 |
| IDA: german | 20 | 700 | 300 | 2 |
| IDA: heart | 13 | 170 | 100 | 2 |
| IDA: image | 18 | 1300 | 1010 | 2 |
| IDA: thyroid* | 5 | 140 | 75 | 2 |
| IDA: titanic | 3 | 150 | 2015 | 2 |
| IDA: twonorm | 20 | 400 | 7000 | 2 |
| IDA: waveform* | 21 | 400 | 4600 | 2 |
| UCI: glass | 9 | 107 | 107 | 6 |
| UCI: wine | 13 | 89 | 89 | 3 |

### A. High-Dimensional Data Sets

First we evaluate the efficiency of EIGML on high-dimensional data in comparison with IGML. The text data set, namely 20newsgroups, consists of posted articles from 20 newsgroups, with roughly 1000 articles per newsgroup. We use the tiny version[3] of this data set with binary occurrence data for 100 words across 16242 postings. The postings have been tagged by the highest level domain in the array "newsgroups". The USPS data set of handwritten digits includes images of "0" through "9" with 1100 examples of each class. The original $28 \times 28$ grayscale images are downsampled to $8 \times 8$ pixels resulting in 64 dimensions. On these two high-dimensional data sets, the Mahalanobis distance is restricted to a set of $k$ basis vectors constructed as follows [16]. We define $\mathbf{R} \in \mathbb{R}^k$ to be the basis of $\mathbf{U}$: $\mathbf{U} = \mathbf{R}(\mathbf{R}^T\mathbf{R})^{-1/2}$. If the number of classes $C$ is greater than $k$, the class-mean points are clustered into $k$ clusters and each cluster's center $\mathbf{r}_i$ is used to form the basis matrix $\mathbf{R} = [\mathbf{r}_1, ..., \mathbf{r}_k]$. If $C \leq k$, we cluster points within each class into approximately $k/C$ clusters and define $\mathbf{R}$ by the center of each cluster. EIGML is also compared with Latent Semantic Analysis (LSA) [26]. LSA projects the data by PCA [23] and computes distances in the projected space. IGML is used as the baseline.



(a) 20newsgroup (b) USPS

Figure 1. Classification error rate (in percent) across bases of varying sizes on (a) 20 newsgroups data set and (b) USPS data set. Compared algorithms are EIGML, LSA [26] and IGML [12].

Figure 1 shows the classification error rate across bases of varying sizes on 20 newsgroups and USPS data sets respectively. Some advantages can be concluded from the results: 1) compared with LSA, EIGML can provide much better performance across all the basis sizes; 2) in terms of quite small bases (e.g., 3 or 4), the error rate of EIGML is higher than the baseline IGML, while EIGML has significantly reduced computational complexity as analysed in Section III; 3) with the bases increasing, EIGML rapidly gets the similar results as IGML. These advantages demonstrate that EIGML is reliable and effective by learning an identity plus low-rank Mahalanobis distance.

### B. IDA and UCI Data Sets

Next we evaluate the effectiveness of SIGML-F and SIGML-L on 13 data sets of varying size and difficulty. IDA data sets are standard for binary classification evaluation [25]. Among them, the ringnorm, twonorm and waveform data sets contain features with only noise. The thyroid and waveform data sets are converted from multi-classes problems so that they contain intrinsic within-class multimodal structures. The banana data set is multimodal as well. SIGML-F is systematically compared with IGML and the following three state-of-the-art metric learning methods: Information-Theoretic Metric Learning (ITML) [5], Large Margin Nearest Neighbor (LMNN) [8], Metric Learning by Collapsing Class (MCML) [4]. The Euclidean distance $d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}(\mathbf{x}_i - \mathbf{x}_j)}$ is used as the baseline. In SIGML-F, the diffusion kernel parameter $\sigma_d$ and the Gaussian width $\sigma$ are tuned in the range $\{0.1, 1, 10\}$ by cross validation.

Figure 2 shows the average error rates on test data sets. Some observations can be drawn from the results. Firstly, SIGML-F consistently outperforms Euclidean distance. In general, the learnt Mahalanobis metrics can result in significantly improved kNN classification accuracy. Secondly, the metric learning methods MCML, ITML and IGML show their limitations for the multimodal data sets with *, since they just exploit the global information. In contrast, despite the Gaussian assumption in SIGML-F, the inclusion of locality notion enables it perform well in multimodal distributions. This observation indicates that the constructed ideal kernel $\mathbf{K}_I$ can well capture the global and the local data structures. Thirdly, SIGML-F shows better metric learning performance than the local algorithm LMNN on most of the data sets. These observations illustrate the generalized and effective performance of SIGML-F for classification.

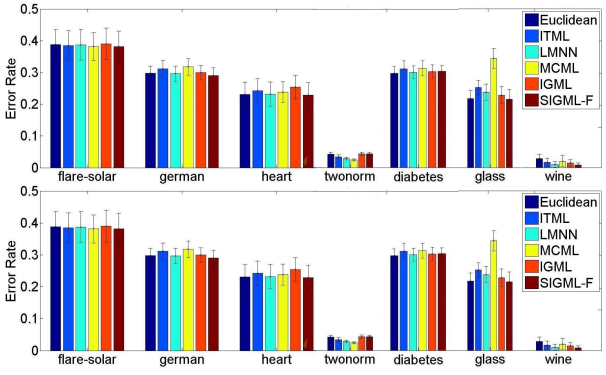



Figure 2. Classification error rate (in percent) on 13 test data sets. Compared algorithms are ITML [5], LMNN [8], MCML [4], IGML [12] and our proposed SIGML-F. Euclidean distance is used as the baseline.

Table II. CLASSIFICATION ERROR RATE (IN PERCENT) AND AVERAGE TRAINING TIME (IN SECONDS) ON WAVEFORM AND TWONORM DATA SETS.

| | | SIGML-L | | | | | | | SIGML-F |
|---|---|---|---|---|---|---|---|---|---|
| Waveform | Rank | 30 | 50 | 100 | 200 | 500 | 800 | 1000 | 3500 |
| | Time | 31.45 | 32.80 | 35.43 | 44.13 | 87.94 | 201.16 | 300.04 | 399.13 |
| | Error Rate | 27.95 | 16.85 | 14.60 | 14.55 | 14.55 | 14.55 | 14.50 | 14.50 |
| Twonorm | Rank | 30 | 50 | 100 | 200 | 500 | 800 | 1000 | 5180 |
| | Time | 66.76 | 71.40 | 91.23 | 94.26 | 122.55 | 231.87 | 339.98 | 1345.13 |
| | Error Rate | 6.41 | 5.68 | 5.65 | 5.65 | 5.59 | 5.56 | 5.55 | 5.55 |

Furthermore, the performance of SIGML-L with the low-rank kernel matrix $\mathbf{K}_{LI}$ is investigated. In this case, waveform and twonorm data sets are employed across different ranks of $\mathbf{K}_{LI}$. Table II shows the error rate on test sets and the running time of SIGML-L. SIGML-F with $\mathbf{K}_I$ is used as the baseline. Unlike previous experiments, the results are obtained by averaging over 10 runs with 70/30 splits of the entire data for training and test sets. The results clearly show that when the rank is quite low (e.g., 30 or 50), SIGML-L greatly decreases the computational time, while the error rate of SIGML-L is close to that of SIGML-F. As the rank becomes high, SIGML-L achieves comparable performance with SIGML-F and benefits from small increase in the running time, especially on twonorm data set. We have similar observations for other data sets, the details are not given due to the lack of space.

## VI. CONCLUSION

In this paper, we have proposed two novel algorithms to alleviate the limitations of Information Geometry Metric Learning (IGML) respectively. The main contribution of this paper is three-fold. (1) By restricting the distance metric to a small dimension basis, we present an algorithm which scales linearly with the dimensionality. Therefore, this proposed method can significantly reduce the time and the space complexity of IGML with competitive performance for high-dimensional data. (2) It is proven that the optimization problem of IGML for low-rank kernels can be cast as a full-rank problem in a lower dimensional space. Based on this theorem, the domain of IGML can be extended to both full-rank and low-rank kernel matrices. (3) We focus on preserving both the local information and the global information in the construction of (full-rank or low-rank) ideal kernel matrices, result in effective and generalized performance. In contrast to most existing metric learning methods, the proposed methods can find closed-form solutions without solving an SDP. Compared with several state-of-the-art metric learning algorithms, extensive experimental results demonstrate the advantages of our methods.

## ACKNOWLEDGEMENT

This work was supported in part by the Natural Science of Foundation of China under Grant 61075051.

## REFERENCES

[1] J. Chen, Z. Zhao, J. Ye, and H. Liu, "Nonlinear adaptive distance metric learning for clustering," in: SIGKDD, 2007, pp. 123-132.

[2] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science, vol. 290, no. 5500, pp. 2323-2326, 2000.

[3] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, "Distance metric learning, with application to clustering with side-information," in: NIPS, 2002, pp. 505-512.

[4] A. Globerson and S. Roweis, "Metric learning by collapsing classes," in: NIPS, 2006, pp. 451-458.

[5] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in: ICML, 2007, pp. 209-216.

[6] R. Jin, S. Wang, and Y. Zhou, "Regularized distance metric learning:theory and algorithm," in: NIPS, 2009, pp. 862-870.

[7] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in: NIPS, 2005, pp. 513-520.

[8] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in: NIPS, 2005, pp. 1473-1480.

[9] L. Yang, R. Jin, R. Sukthankar, and Y. Liu, "An efficient algorithm for local distance metric learning," in: AAAI, 2006, pp. 543-548.

[10] G. Q. Zhong, K. Z. Huang, and C. L. Liu, "Low rank metric learning with manifold regularization," in: ICDM, 2011, pp. 1266-1271.

[11] A. Bar-Hillel, T. Hertz, N. Shental, and D. Weinshall, "Learning a Mahalanobis metric from equivalence constraints," J. Mac. Learn. Res., vol. 6, pp. 937-965, 2005.

[12] S. Wang and R. Jin, "An information geometry approach for distance metric learning," in: AISTATS, 2009, pp. 591-598.

[13] N. Cristianini, J. Kandola, A. Elisseeff, and J. Shawe-Taylor, "On kernel target alignment," in: NIPS, 2002, pp. 367-373.

[14] K. Tsuda, S. Akaho, K. Asai, and C. Williams, "The EM algorithm for kernel matrix completion with auxiliary data," J. Mac. Learn. Res., vol. 4, pp. 67-81, 2003.

[15] L. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," USSR Comp. Mathematics and Mathematical Physics, vol. 7, pp. 200-217, 1967.

[16] P. Jain, B. Kulis, J. V. Davis, and I. S. Dhillon, "Metric and kernel learning using a linear transformation," J. Mac. Learn. Res., vol. 13, pp. 519-547, 2012.

[17] J. V. Davis and I. S. Dhillon, "Structured metric learning for high dimensional problems," in: KDD, 2008, pp. 195-203.

[18] D. A. Harville, *Matrix Algebra From a Statistician's Perspective*. Springer Berlin, 2008.

[19] R. S. Kondor and J. Lafferty, "Diffusion kernels on graphs and other discrete input spaces," in: ICML, 2002, pp. 315-322.

[20] A. Smola and R. Kondor, "Kernels and regularization on graphs," in: COLT, 2003, pp. 144-158.

[21] B. Kulis, A. Surendran, and J. Platt, "Fast low-rank semidefinite programming for embedding and clustering," in: AISTATS, 2007.

[22] B. Kulis, M. A. Sustik, and I. S. Dhillon, "Low-rank kernel learning with bregman matrix divergences," J. Mac. Learn. Res., vol. 10, pp. 341-376, 2009.

[23] I. Jolliffe, *Principal Component Analysis*. Springer-Verlag, 1986.

[24] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, Cambridge, 2004.

[25] G. Rëtsch, T. Onoda, and K.-R. Müller, "Soft margins for adaboost," Mach. Learn., vol. 42, pp. 287-320, 2001.

[26] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by latent semantic analysis," J. Am. Soc. Inf. Sci. Tec., vol. 41, pp. 391-407, 1990.