An Efficient Sparse Metric Learning in High-Dimensional Space via ℓ_1-Penalized Log-Determinant Regularization

Guo-Jun Qi QI4@ILLINOIS.EDU
Depart. ECE, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, IL 61801 USA

Jinhui Tang, Zheng-Jun Zha, Tat-Seng Chua {TANGJH, ZHAZJ, CHUATS}@COMP.NUS.EDU.SG
School of Computing, National University of Singapore, Computing 1, 13 Computing Drive, Singapore 117417

Hong-Jiang Zhang HJZHANG@MICROSOFT.COM
Microsoft Advanced Technology Center, 49 Zhichun Road, Beijing 100190 China

Abstract

This paper proposes an efficient sparse metric learning algorithm in high dimensional space via an ℓ_1-penalized log-determinant regularization. Compare to the most existing distance metric learning algorithms, the proposed algorithm exploits the sparsity nature underlying the intrinsic high dimensional feature space. This sparsity prior of learning distance metric serves to regularize the complexity of the distance model especially in the "less example number p and high dimension d" setting. Theoretically, by analogy to the covariance estimation problem, we find the proposed distance learning algorithm has a consistent result at rate $\mathcal{O}\left(\sqrt{(m^2 \log d)/n}\right)$ to the target distance matrix with at most m nonzeros per row. Moreover, from the implementation perspective, this ℓ_1-penalized log-determinant formulation can be efficiently optimized in a block coordinate descent fashion which is much faster than the standard semi-definite programming which has been widely adopted in many other advanced distance learning algorithms. We compare this algorithm with other state-of-the-art ones on various datasets and competitive results are obtained.

1. Introduction

An appropriate distance metric has become a fundamental tool in many supervised and unsupervised learning algorithms, such as the nearest neighborhood classification, kernel methods, k-means and many others. Moreover, in a variety of applications such as image retrieval and indexing (Hoi et al., 2008), a proper distance metric plays an important role to measure the relevance/irrelevance between different images. Therefore, to apply a proper distance into these practical uses, many distance learning algorithms have been proposed to reveal the intrinsic metric revealing the semantic meaning between different samples.

Although many existing algorithms for metric learning have been shown to perform well in various applications, most of them do not explicitly deal with learning of a distance metric in a higher dimensional input space with smaller sample size. This high-dimensional problem exists in a wide range of applications from image retrieval, face recognition and computational biology to natural language processing, where the dimension of feature space may be comparable to or substantially larger than the sample size. It is well-known that such "curse of dimensionality" problem leads to serious breakdown in many algorithms with an under-determined problem. In the absence of additional model assumptions, it is not well conditioned to obtain consistent procedures when the feature dimension is much larger compared to the sample size. An effective way to overcome this problem is to impose some restrictions or prior knowledge information on the model, which regularizes the model complexity so that it only requires a smaller number of examples to learn a well posed metric from the perspective of machine learning theory.

Recently, Euclidean metric prior has been widely used in many metric learning literatures, such as (Davis et al., 2007), (Schultz & Joachims, 2004) and (Xing et al., 2003). It imposes a prior on the metric to be as close as the Euclidean distance. For example, (Davis

Appearing in *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada, 2009.

et al., 2007) employs an information-theoretic regularization term to respect the Euclidean distance in the input feature space by minimizing the Bregman divergence between the Mahalanobis distance matrix and the identity matrix corresponding to the Euclidean distance, subject to a set of linear constraints. By minimizing Bregman divergence, the learned Mahalanobis matrix usually tends to be as close as the identity matrix. Since the off-diagonal elements of the identity matrix are all zeros, the identity matrix itself is rather sparse. But minimizing the Bregman divergence from the identity matrix cannot guarantee the obtained Mahalanobis matrix to be sparse as well. We will show a formulation of a sparse Mahalanobis matrix reflects the intrinsic nature of sparsity underlying the input (especially high dimensional) space from both practical and theoretical perspectives that serves as the key prior in the proposed metric learning algorithm.

We motivate the proposed distance metric learning algorithm from the following three aspects.

I We impose a sparse prior on the off-diagonal elments of Mahalanobis matrix to learn a compact distance metric in a high dimensional space. This prior can be justified from three aspects. First, from a practical perspective, a sparse Mahalanobis matrix with only a small portion of off-diagonal elements nonzero complies with the fact that in high dimensional input spaces, the off-diagonal elements of concentration matrix (i.e. the inverse of the covariance matrix)are often remarkably small which can be safely ignored. This observation reflects the sparse correlations between different dimensions and supports a sparse Mahalanobis metric. Second, analogous to the principle of minimum description length (Hansen & Yu, 2001) in model selection, the sparse principle that yields the most compact model is preferred in distance metric learning. Third, a sparse Mahalanobis distance can be computed very efficiently which is of significant importance to many realistic applications.

II By analogous to covariance estimation (Ravikumar et al., 2008), the proposed sparse metric learning algorithm in Section 3 has a consistent distance estimation at rate $||\widetilde{M} - M^\star||_F = \mathcal{O}\left(\sqrt{(m^2 \log d)/n}\right)$, with high probability, for the concentration matrix with at most m nonzero per row, where $\widetilde{M}$ and $M^\star$ are the estimated and target distance matrix respectively, d is the dimension number and n is the sample size. This rate reveals a much consistent result even in the "large d, small n" setting as long as the number of nonzero elements m is small enough(i.e., the sparsity requirement).

III We show that the obtained ℓ_1-penalized Log-Determinant optimization problem for the sparse metric can be efficiently minimized by leveraging a block coordinate descent fashion algorithm (Friedman et al., 2007), which is much faster than the Semi-Definite Programming (SDP) methods widely used in metric learning.

The remainder of this paper is organized as follows. In section 2, we first motivate the proposed ℓ_1-penalized log-determinant regularization framework. We explain the superiority of sparsity property from both practical and theoretical perspectives. Section 3 details the proposed sparse distance metric learning algorithm and an efficient block coordinate descent optimization method for this formulation. Section 4 evaluates the proposed distance metric learning algorithm by comparison with the other state-of-the-art algorithms. Finally we conclude the paper in Section 5.

2. ℓ_1-Penalized Log-Determinant Regularization

We begin this section with some notational definitions. Our goal is to learn a (squared) Mahalanobis distance

$$d_M(x, y) = (x - y)^T M (x - y) \tag{1}$$

from a set of n examples $\{x_1, x_2, \cdots, x_n\}$ in an input feature space $\mathbb{R}^d$, where M is a $d \times d$ positive semi-definite matrix. Two sets of pairwise similarity constraints $\mathcal{S} = \{(x_i, x_j) | x_i \text{ and } x_j \text{ are similar}\}$ and dissimilarity constraints $\mathcal{D} = \{(x_i, x_j) | x_i \text{ and } x_j \text{ are dissimilar}\}$ are also given. The purpose is to learn a Mahalanobis distance (1) by leveraging the above similarity and dissimilarity constraints on these n examples.

We start by imposing a suitable prior knowledge on the Mahalanobis matrix for learning. As stated in Section 1, (Davis et al., 2007) biject the Mahalanobis distance to an equal-mean multivariate Gaussian distribution and formulate the problem by minimizing the differential relative entropy between the two multivariate Gaussians as its prior information. In detail, given a Mahalanobis matrix M, the corresponding multivariate Gaussian distribution can be expressed as $p(x; M) = \frac{1}{Z} \exp\left(-\frac{1}{2}(x - \mu)^T M (x - \mu)\right)$, where Z is a normalizing constant and M is the concentration matrix of the distribution. In other words, they biject Mahalanobis matrix into a corresponding concentra-

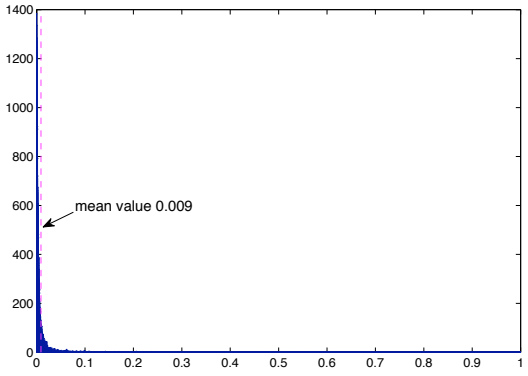

Figure 1. It illustrates the distribution of absolute values of the elements of concentration matrix Σ^{-1} over a set of 225 dimensional feature vectors. These values are scaled into $[0, 1]$. The figure illustrates most of these values are much closer to zero and only a small portion have large values. Also, 81.60% of elements are less than mean value 0.009.

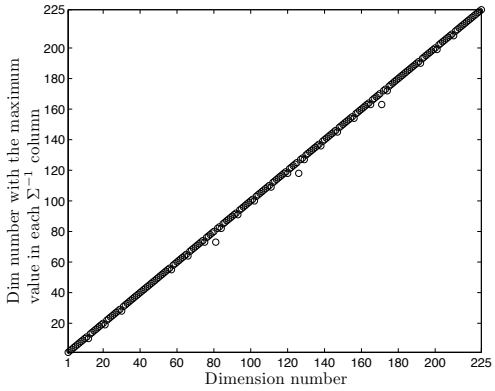

Figure 2. From this figure we can find the diagonal elements have the largest values among the elements corresponding to each dimension in the concentration matrix. Of all the 225 dimensions, the diagonal elements have the largest values on 202 dimensions (about 89.78%).

tion matrix for the multivariate Gaussian. This multivariate Gaussian distribution reflects the distribution of samples revealed by the learned Mahalanobis distance. However, a simple bijection between Mahalanobis distance and equal-mean multivariate Gaussian distribution oversimplifies the underlying metric structure in (Davis et al., 2007). For example, if we require the learned metric to be as close as the Euclidean distance corresponding to the identity matrix as its Mahalanobis matrix, no existing practice and theory can guarantee that the complexity of the resultant metric has been compactly regularized especially in the "less sample, high dimensions" setting.

In this paper, we impose a different prior on learning Mahalanobis distance. The observation is the sparsity of sample concentration matrix Σ^{-1}, where Σ is the covariance matrix of examples. Figure 1 illustrates a distribution of the absolute values of the elements in Σ^{-1}, which are calculated from a set of 225 dimensional features vectors. These values have been normalized into the region $[0, 1]$. It illustrates that most of the elements have a much smaller values close to zero and only a small portion has a larger values. According to the above mentioned bijection between the Mahalanobis distance and the multivariate Gaussian in (Davis et al., 2007), it shows that a sparse Mahalanobis matrix is preferred. This sparsity nature results from the weak correlation among different dimensions in the high dimensional space because most of the different features are measured by different mechanisms in the real world. For example, in image retrieval we usually extract features from different color components, such as in CIE Luv color space. These components are generated by nearly independent lighting components which have very week correlations between each other. On the other hand, the diagonal elements often have the largest values compared to those off-diagonal elements as illustrated in Figure 2, and they are much larger than those off-diagonal elements (see Figure 3) due to the the strong self correlations. In summary, we can obtain a sparsity prior where the off-diagonal elements of Mahalanobis matrix have much small values closer to zero as compared to those diagonal elements. Here we verify this sparsity prior from a practical viewpoint. In Section 3, we will show that a more consistent result can be obtained by applying the sparsity prior to metric learning with the proposed l_1-penalized log-determinant formulation below from a theoretical perspective.

To seek a sparse solution for Mahalanobis distance (1), we can formulate to minimize l_0-norm of Mahalanobis matrix M which counts its nonzero elements. However, the problem for l_0-minimization is that it is NP-hard and difficult even to approximate (Wright et al., 2008). Instead, if the target distance matrix M is sparse enough, we can equivalently minimize ℓ_1-norm of M, i.e., the absolute summation of all its elements instead (Donoho, 2006). We also note that those diagonal elements of M are often not so sparse as indicated in Figure 2 and 3 where the diagonal elements usually have much more significant values than those off-diagonal elements. Therefore we propose to mini-

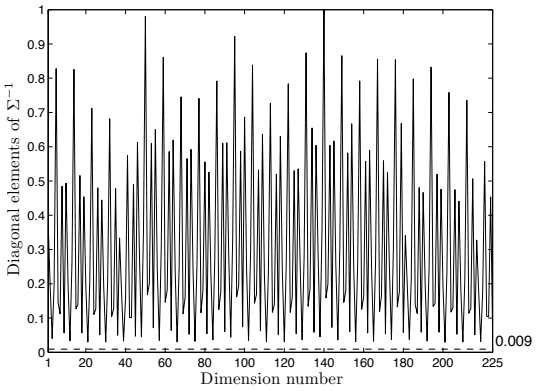

Figure 3. Self variances of different dimensions. We find that as compared to the other portions of the elements as illustrated in Figure 1, all the diagonal elements are larger than the mean value 0.009.

mize the off-diagonal ℓ_1-norm

$$||M||_{1,\text{off}} = \sum_{i \neq j} |M_{ij}| \tag{2}$$

to pursue a sparse solution.

In addition to the above sparsity requirement, we also impose another *distribution* prior on the metric learning that parameterizes the Mahalanobis distance by the sample concentration or the identity matrix corresponding to the squared Euclidean distance. The sample concentration prior provides the metric learning with the distribution knowledge about the sample distribution, while the identity matrix gives the most unbiased prior to learn a metric starting from the squared Euclidean distance. In other words, we regularize the Mahalanobis matrix M to be as close as possible to a given Mahalanobis matrix M_0 associated with the priori sample distribution by minimizing the log-determinant divergence $D_g(M||M_0) = g(M) - g(M_0) - \langle \nabla_g(M_0), M - M_0 \rangle$ between M and M_0 with a strict convex, continuously differentiable function $g(M) = -\log\det(M)$. This log-determinant divergence function can be rewritten as

$$D_g(M||M_0) = \text{tr}\left(M_0^{-1}M\right) - \log\det M \tag{3}$$

where $\text{tr}(\cdot)$ means the trace operation on matrix. Note that we ignore the constant term regarding M_0 in the above equation.

The above sparsity and distribution priors complement each other. The sparsity prior prefers a sparse metric structure in the general sense to control its complexity while the distribution prior prefers a metric model as close as possible to the metric calculated from the priori sample distribution in the data-driven sense. By combining the sparsity prior and log-determinant regularization together, we can obtain the following sparse metric learning formulation

$$\begin{aligned} &\min_{M} \ \text{tr}\left(M_0^{-1}M\right) - \log\det M + \lambda||M||_{1,\text{off}} + \eta\mathcal{L}(\mathcal{S},\mathcal{D}) \\ &s.t. \ \ M \succeq 0 \end{aligned} \tag{4}$$

where $M \succeq 0$ is the positive semi-definite constraint, λ is the balance parameter trading off between sparsity prior and the M_0 prior, and $\mathcal{L}(\mathcal{S},\mathcal{D})$ is a loss function defined on the sets of similarity $\mathcal{S}$ and dissimilarity $\mathcal{D}$ constraints. η is a positive balance parameter trading off between the loss function and the regularizer. We will detail it in the following Section.

3. Sparse Distance Metric Learning

In this section, we first complete the formulation (4) by designing a proper loss function $\mathcal{L}(\mathcal{S},\mathcal{D})$. Then, we propose to use an efficient optimization algorithm to solve ℓ_1-penalized log-determinant problem in a block coordinate descent fashion.

3.1. Formulation

Given the sets of similar and dissimilar pairs $\mathcal{S}$ and $\mathcal{D}$, we can define an incidence matrix K to encode the (dis)similarity information as

$$K_{ij} = \begin{cases} 1, \text{ if } (x_i,x_j) \in \mathcal{S} \\ -1, \text{ if } (x_i,x_j) \in \mathcal{D} \end{cases} \tag{5}$$

Like (Hoi et al., 2008)(Zhang et al., 2007), we can assume that there exists a corresponding linear transformation $A : \mathbb{R}^m \rightarrow \mathbb{R}^l$, where $A \in \mathbb{R}^{m\times l}$ and $M = AA^T$ parameterizes the Mahalanobis matrix so that the squared Euclidean distance in the transformed space $\mathbb{R}^l$ can be computed as

$$\begin{aligned} &d_M(x_i,x_j) \\ &= ||A^Tx_i - A^Tx_j||_2^2 \\ &= (x_i-x_j)^T AA^T (x_i-x_j) \\ &= (x_i-x_j)^T M (x_i-x_j) \end{aligned} \tag{6}$$

By minimizing the distances between those adjacent examples indicated in K, we can formulate to minimize

the following loss function

$$\begin{aligned}
&\mathcal{L}(\mathcal{S},\mathcal{D}) \\
&= \tfrac{1}{2}\sum_{i,j=1}^{n} ||A^T x_i - A^T x_j||_2^2 K_{ij} \\
&= \sum_{i,j=1}^{n} \left(x_i^T A A^T x_i - x_i^T A A^T x_j\right) K_{ij} \\
&= \sum_{i,j=1}^{n} \left(x_i^T M x_i - x_i^T M x_j\right) K_{ij} \\
&= \mathrm{tr}\left(X^T M X D\right) - \mathrm{tr}\left(X^T M X K\right) \\
&= \mathrm{tr}\left(X D X^T M\right) - \mathrm{tr}\left(X K X^T M\right) \\
&= \mathrm{tr}\left(X\left(D-K\right)X^T M\right) \\
&= \mathrm{tr}\left(X L X^T M\right)
\end{aligned} \tag{7}$$

where $X = [x_1, x_2, \cdots, x_n]$, D is a diagonal matrix whose diagonal elements are the sums of the row elements of K, $L = D - K$ is the Laplacian matrix.

Substitute Eqn. (7) into the formulation (4), we can obtain the following ℓ_1-penalized log-det optimization problem

$$\begin{aligned}
&\min_{M}\ \mathrm{tr}\left(M_0^{-1} M\right) - \log\det M + \lambda||M||_{1,\text{off}} + \eta\mathcal{L}(S,D) \\
&= \mathrm{tr}\left(\left(M_0^{-1} + \eta X L X^T\right)\cdot M\right) - \log\det M + \lambda||M||_{1,\text{off}} \\
&s.t.\ \ M \succeq 0
\end{aligned} \tag{8}$$

Note that since both $-\log\det(M)$ and $||M||_{1,\text{off}}$ are convex w.r.t. M, and $\mathrm{tr}\left(\left(\mathrm{M_0^{-1}} + \eta \mathrm{X L X^T}\right)\cdot \mathrm{M}\right)$ is a linear term, the above formulation is convex and there exists a global minimum in the above optimization problem.

Denote $P = M_0^{-1} + \eta X L X^T$, we analogize the above metric learning problem to the covariance estimation problem with a "pseudo" sample covariance matrix P, which plays the same role as the sample covariance matrix underlying the consistency analysis in (Ravikumar et al., 2008). Here we express their main result about consistent analysis. For a target Mahalanobis matrix with at most m nonzero per row subject to sub-Gaussian condition, minimizing an ℓ_1-penalized log-determinant function in Eqn.(8) leads to a consistent distance estimate at rate $||\widetilde{M} - M^\star||_F = \mathcal{O}\left(\sqrt{\left(m^2\log d\right)/n}\right)$, with high probability, where $\widetilde{M}$ and $M^\star$ are the estimated and target distance matrix respectively, d is the dimension number and n is the sample size. The rate reveals a much consistent result even in the "large dimension number d, small training size n" setting as long as the number of nonzero elements per row is small enough per row. It justifies the sparsity prior to learn a metric distance from a theoretical perspective.

The above semi-definite optimization problem (8) contains a ℓ_1-penalized log-det operator in the objective function which can be converted into an associated standard Semi-Definite Programming (SDP) problem (Todd, 2001) by making the following translations $M = U - V$, where $U \geq 0$ and $V \geq 0$. If either U_{ij} or V_{ij} has to be zero, we have $||M||_{1,\text{off}} = ||U||_{1,\text{off}} + ||V||_{1,\text{off}} = \sum_{i\neq j} U_{ij} + \sum_{i\neq j} V_{ij}$. Therefore, the SDP problem (8) can be formulated as

$$\begin{aligned}
&\min_{U,V}\ \mathrm{tr}\left(\left(M_0^{-1} + \eta X L X^T\right)\cdot(U-V)\right) - \log\det(U-V) \\
&+\ \ \lambda||U||_{1,\text{off}} + \lambda||V||_{1,\text{off}} \\
&=\ \ \mathrm{tr}\left(\left(M_0^{-1} + L\right)(U-V)\right) - \log\det(U-V) \\
&+\ \ \lambda\sum_{i,j=1,i\neq j}^{d} U_{ij} + \lambda\sum_{i,j=1,i\neq j}^{d} V_{ij} \\
&s.t.\ U - V \succeq 0 \\
&\quad U \geq 0, V \geq 0
\end{aligned} \tag{9}$$

To validate the equivalence between problem (8) and (9), we have the following lemma

Lemma 1. *In the problem (9), either U_{ij} or V_{ij} has to be zero so that $||M||_{1,\text{off}} = ||U||_{1,\text{off}} + ||V||_{1,\text{off}} = \sum_{i\neq j} U_{ij} + \sum_{i\neq j} V_{ij}$.*

Proof. If this theorem does not hold, we can assume $U_{ij} \geq V_{ij} > 0$ without any loss of generality. Then we can find a better solution by setting $U_{ij} \leftarrow U_{ij} - V_{ij}, V_{ij} \leftarrow 0$. It can be easily proved that this new solution still satisfies all the above constraints, and the new objective value is smaller than before. This contradicts the optimality of U, V, thus the we can always find a more optimal solution of U_{ij}, V_{ij}, either of which is zero. □

With the above lemma, it is not difficult to prove that the SDP problem (9) yields solutions equivalent to those obtained by ℓ_1- minimization problem (8).

3.2. An Efficient ℓ_1-penalized Log-Determinant Solver

Although there exist optimization algorithms that directly solves the semi-definite problem like (8), their computational costs are usually expensive especially when the feature dimension is high. An alternative optimization method is to use the block coordinate descent algorithm (Friedman et al., 2007). Slightly different from (8), they optimize over the inverse of the Mahalanobis matrix M^{-1} rather than M directly. Let W be an estimation of M^{-1}, they show that the problem can be optimized over each row and the corresponding column of W in a block coordinate descent fashion. With a partitioning W and $P = M_0^{-1} + \eta X L X^T$

$$W = \begin{bmatrix} W_{11} & w_{12} \\ w_{12}^T & w_{22} \end{bmatrix} \tag{10}$$

and

$$P = \begin{bmatrix} P_{11} & p_{12} \\ p_{12}^T & p_{22} \end{bmatrix} \tag{11}$$

(Banerjee et al., 2007) proves the solution for w_{12} satisfies

$$w_{12} = \arg\min_z \left\{ z^T W_{11}^{-1} z : ||z - p_{12}||_\infty \leq \lambda \right\} \tag{12}$$

By permuting the rows and columns so the target column is always placed the last, problem (12) is solved for each column so as to update the estimates of W after each step until convergence. With an initial positive definite matrix, it can be shown that the iterations from this procedure remain positive definite even if $p > n$. Problem (12) can be solved by its associated dual problem (Banerjee et al., 2007)

$$\min_\alpha \left\{ \frac{1}{2} ||W_{11}^{1/2}\alpha - W_{11}^{-1/2} p_{12}||^2 + \lambda ||\alpha||_1 \right\} \tag{13}$$

If α solves it, $w_{12} = W_{11}\alpha$ solves (12). The above lasso problem can be efficiently solved by a coordinate descent algorithm. Let $T = W_{11}$ and $g = p_{12}$, α can be updated until convergence as

$$\hat{\alpha}_j \leftarrow \frac{F\left(g_j - \sum_{k \neq j} T_{kj} \hat{\alpha}_k, \lambda \right)}{T_{jj}} \tag{14}$$

where $F(x,t) = \text{sign}(|\text{x}| - \text{t})_+$ is the soft-threshold operator. Finally, an estimation W is obtained and the Mahalanobis matrix can be recovered by $M = W^{-1} = \begin{bmatrix} M_{11} & m_{12} \\ m_{12}^T & m_{22} \end{bmatrix}$ as

$$\begin{aligned} m_{12} &= -\hat{\alpha} m_{22} \\ m_{22} &= 1/\left(w_{22} - w_{12}^T \hat{\alpha} \right) \end{aligned} \tag{15}$$

in the final step through the corresponding α for each row. Finally, we summarize the above block coordinate descent method in Algorithm 1.

Algorithm 1 Optimization Algorithm for (8)

input example matrix $X = [x_1, x_2, \cdots, x_n]$, Laplacian matrix L, balance parameters λ, η.
  Set matrix $P = M_0^{-1} + \eta X L X^T$.
  Initialize $W = P + \lambda I$ and the diagonal elements of W remain changed in the following.
  repeat
    Solve the problem (13):
    repeat
      for $j = 1$ **to** $d - 1$ **do**
        update $\hat{\alpha}_j \leftarrow \frac{F(g_j - \sum_{k \neq j} T_{kj}\hat{\alpha}_k, \lambda)}{T_{jj}}$
      end for
    until convergence of $\hat{\alpha}$
    Fill in the corresponding row and column of W using $w_{12} = W_{11}\hat{\alpha}$.
    Permute the rows and columns of W for the next target w_{12}.
  until convergence of W
  Compute the Mahalonobis matrix M:
  for each α of the corresponding row of M **do**
    $m_{12} = -\hat{\alpha} m_{22}$
    $m_{22} = 1/\left(w_{22} - w_{12}^T \hat{\alpha} \right)$
    Fill in the corresponding row and column of M using m_{12}, and the corresponding diagonal element using m_{22}.
  end for
output The Mahalonobis matrix M.

4. Experiments

In this Section, we compare the proposed Sparse Distance Metric Learning (SDML) with other existing state-of-the-art algorithms on various benchmark UCI datasets and a real image dataset. We compare these algorithms from the following three aspects

1. k-nearest neighbor (k-NN) classification performance by using the different distance metrics.

2. The changes of k-NN classification performances under different ratios of n (i.e., the example size) to d (i.e., the dimension number). It illustrates how these metrics perform in the "less n, large d" settings.

3. We also compare the computational costs of these algorithms on the same platform.

We evaluate the algorithm on four UCI benchmark datasets - Iris, Ionosphere, Wine, Sonar and a real-world image dataset Corel. All the metrics are compared via two-fold cross validation with $k = 3$ for k-NN classification.

4.1. k-NN Classification Results

The proposed SDML is compared with the following algorithms for k-NN classification

Euclidean The squared Euclidean distance as a baseline algorithm.

InvCov a Mahalanobis distance parameterized by the inverse of the sample covariance. It is equivalent to performing a standard PCA transformation over the input space followed by computing

the squared Euclidean distance in the transformed space.

LMNN Large margin nearest neighbor proposed by (Weinberger et al., 2005). It is trained with the goal that the k-nearest neighbors always belong to the same class while examples from different classes are separated by a large margin.

ITML Information-theoretic metric learning proposed by (Davis et al., 2007). It formulates to learn the Mahalanobis distance by minimizing the differential relative entropy between two multivariate Gaussians under a set of constraints on the distance function.

For the proposed SDML, we use two different Mahalanobis matrices as prior matrix M_0. They are the identity matrix which regularizes to the squared Euclidean distance, and sample covariance matrix calculated on the training examples. We denote them by SDML(Identity Matrix) and SDML(Covariance Matrix) respectively.

The experimental results are illustrated in Table 1. We can see that the SDML has incurred the smallest error rates across all the datasets compared to the other distance metrics. On the other hand, the SDML(Covariance Matrix) performs better than SDML(Identity Matrix). This is probably due to the underlying sample distribution revealed by the sample covariance.

4.2. Performance Changes under different n/d

We also conduct experiments to illustrate how the proposed SDML performs with less training examples n and higher dimensions d. The experiments are done on k-NN classification with $k = 3$ on an image dataset Corel. This dataset contains 500 images from the five different classes each of which contains 100 images. We split the whole dataset into three parts. The first is the training set with 250 images, the second is the validation set with 100 images and the last is the testing set with 150 images. A 225 dimensional color moment feature vector is extracted from each individual image. In experiments, we use the varying numbers of the training examples to learn the distance metric and compare the performance changes under different n/d, where a lower n/d means a relatively high dimensional feature space compared to a smaller number of training examples.

Figure 4 illustrates SDML(Identity Matrix) significantly outperforms the other metrics under the different n/d ranging from 0.09 to 0.27, especially when n/d is low. It empirically verifies that SDML can regularize the metric model in a relatively high feature space resulting in the much competitive performance.

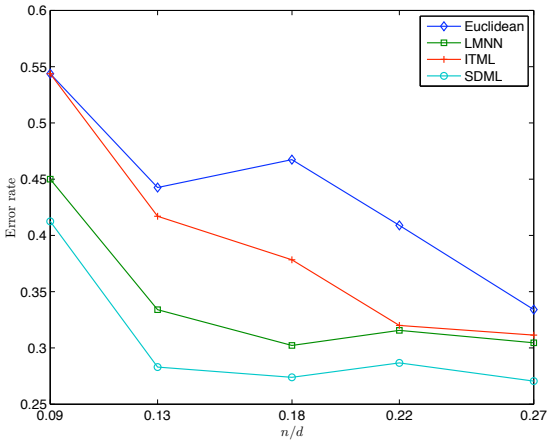

Figure 4. Comparison of k-NN classification error rates with different distance metrics by the changes of n/d. It shows that the proposed SDML performs best compared to the other metrics especially when n/d is low (i.e., the dimension is relatively high compared to the number of training examples.

Table 2. Training time used for the different distance learning algorithm on the different data sets.

ALGORITHM	LMNN	ITML	SDML
IRIS	1.230	0.01545	0.0071
IONOSPHERE	6.440	0.02087	0.0083
WINE	2.82	0.3945	0.058
SONAR	10.46	1.3598	0.098

4.3. Computational Costs

Finally, we compare the computational efficiency of learning these distance metrics. All these experiments are conducted on a PC equipped with an Intel 2.66 GHz CPU and 3.25 GMB memory.

In Table 2, we report the training time used to learn these metrics over the benchmark datasets, and Figure 5 illustrates the training time on Corel image dataset under different n/d. We find that SDML is two and three orders of magnitude faster than LMNN and ITML, respectively.

Table 1. k-NN classification error rates (%) for the different distances across various benchmark datasets. We can see the SDML gains the best performance across all the used datasets.

Algorithm	Iris	Ionosphere	Wine	Sonar
Euclidean	4.00	14.86	4.50	18.27
InvCov	8.67	17.71	43.82	39.42
LMNN	3.34	14.29	2.25	14.42
ITML	3.00	17.14	3.94	23.56
SDML(Identity Matrix)	**2.00**	13.71	0.5618	16.35
SDML(Inverse Covariance)	**2.00**	**12.00**	**0**	**13.46**

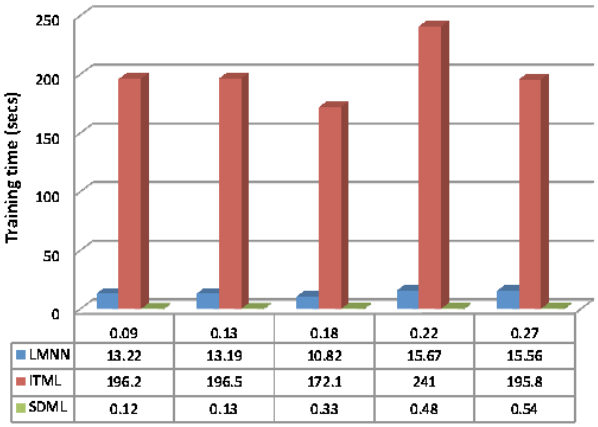

Figure 5. Comparison of the computational times used to learn the distance metric by the changes of n/d. It shows that SDML is much faster as compared to LMNN and ITML.

5. Conclusion

This paper presents an efficient sparse metric learning algorithm via ℓ_1-penalized log-determinant regularization. This *sparsity* prior on the distance metric regularizes the complexity of the distance model especially with the "less example number p and high dimensions d" in a general sense. On the other hand, we propose a complementary *distribution* prior which prefers a model as close as possible to a metric calculated from the priori sample distribution in the data-driven sense. We show that the ℓ_1-penalized log-determinant function can be efficiently optimized by a block coordinate descent algorithm. The experiments on both benchmark datasets and the image dataset show that competitive results can be obtained as compared to the other state-of-the-art algorithms.

References

Banerjee, O., Ghaoui, L. E., & d'Aspremont, A. (2007). Model selection through sparse maximum likelihood estimation. *Journal of Machine Learning Research*.

Davis, J. V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. *Proc. of ICML*.

Donoho, D. (2006). For most large underdetermined systems of linear equations the minimum l1-norm is also the sparsest solution. *Comm. On Pure and Applied Math, 59*, 797–829.

Friedman, J., Hastie, T., & Tibshirani, R. (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostat*.

Hansen, M., & Yu, B. (2001). Model selection and the minimum description length principle. *Journal of the American Statistical Association, 96*, 746–774.

Hoi, S. C. H., Liu, W., & Chang, S.-F. (2008). Semi-supervised distance learning for collaborative image retrieval. *Proc. of IEEE CVPR*.

Ravikumar, P., Wainwright, M. J., Raskutti, G., & Yu, B. (2008). *High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence* (Technical Report 767). Department of Statistics, University of California, Berkeley.

Schultz, M., & Joachims, T. (2004). Learning a distance metric from relative comparisons. *Proc. of NIPS*.

Todd, M. (2001). Semidefinite optimization. *Acta Numerica*, 515–560.

Weinberger, K. Q., Blitzer, J., & Saul, L. K. (2005). Distance metric learning for large margin nearest neighbor classification. *Proc. of NIPS*.

Wright, J., Yang, A. Y., Ganesh, A., Sastry, S. S., & Ma, Y. (2008). Robust face recognition via sparse representation. *IEEE Transactions on Pattern Recognition and Machine Intelligence*.

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (2003). Distance metric learning, with application to clustering with side-information. *Proc. of NIPS*.

Zhang, W., Xue, X., Sun, Z., Guo, Y.-F., & Lu, H. (2007). Optimal dimensionality of metric space for classification. *Proc. of ICML*.