Distance Metric Learning by Optimization on the Stiefel Manifold

Ankita Shukla
Saket Anand
IIIT Delhi, New Delhi, 110020.

Abstract
Distance metric learning has proven to be very successful in various problem domains. Most techniques learn a global metric in the form of a \( n \times n \) symmetric positive semidefinite (PSD) Mahalanobis distance matrix, which has \( O(n^2) \) unknowns. The PSD constraint makes solving the metric learning problem even harder making it computationally intractable for high dimensions. In this work, we propose a flexible formulation that can employ different regularization functions, while implicitly maintaining the positive semidefiniteness constraint. We achieve this by eigendecomposition of the rank \( p \) Mahalanobis distance matrix followed by a joint optimization on the Stiefel manifold \( S_{n,p} \) and the positive orthant \( \mathbb{R}^p_{+} \). The resulting non-convex optimization problem is solved by employing an alternating strategy. We use a recently proposed projection free approach for efficient optimization over the Stiefel manifold. Even though the problem is nonconvex, we empirically show competitive classification accuracy on UCI and USPS digits datasets.

1. Introduction
Distance metric learning has received a lot of attention in the last decade owing to its success in many application domains like computer vision, classification and clustering. The default Euclidean distance equally weights each dimension in the input space and is often inadequate to capture the semantics of the data. Metric learning techniques use training examples to learn a distance function that is semantically consistent with the data. The commonly used Mahalanobis distance metric is characterized by a positive semidefinite (PSD) matrix that applies a linear transformation in the input space.

Many popular techniques (Davis et al., 2007; Kulis et al., 2009; Jain et al., 2012; Law et al., 2014) set up the metric learning problem in a constrained optimization framework. The imposed constraints capture the intuition that same class point pairs have small distances, while sample points from different classes have a large distance. The challenge in solving such problems is efficient projection on to the constraint space while maintaining the positive semidefiniteness of the Mahalanobis distance matrix.

Techniques that rely on projection on to \( S_{n}^{+} \) like (Weinberger & Saul, 2009; Xing et al., 2002; Law et al., 2014), usually require an eigen-decomposition or SVD in each iteration resulting in an additional cost of \( O(n^3) \). Projection free approaches like (Davis et al., 2007; Kulis et al., 2009; Jain et al., 2012) use special regularization functions leading to updates that guarantee positive semidefiniteness. In this paper, we explore a projection free approach that permits the flexibility to use different regularization functions.

The remainder of the paper is organized as follows. We provide a brief review of Stiefel manifold in Section 2 followed by the details of the proposed framework in Section 3. The experimental results are presented in Section 4. We discuss the extension of our parametrization in Section 5 and conclude in Section 6.

2. Optimization on the Stiefel Manifold
The set of \( n \times p \) orthonormal matrices has a Riemannian structure and is called the Stiefel manifold, \( S_{n,p} = \{ U \in \mathbb{R}^{n \times p} : U^\top U = I_p, n \geq p \} \) (Edelman et al., 1998). An alternate interpretation is that of a quotient space of the orthogonal group \( O_n = \{ Q \in \mathbb{R}^{n \times n} : Q^\top Q = I_n \}, i.e., S_{n,p} = O_n/O_{n-p} \). The tangent space at a point \( U \in S_{n,p} \) is given by \( T_U = \{ \Delta \in \mathbb{R}^{n \times p} : \Delta^\top U = -U^\top \Delta \} \).
Wen and Jin (Wen & Yin, 2013) proposed an efficient constraint preserving update on the Stiefel manifold based on the Cayley transformation. The key idea is to relax the constraint of moving along geodesics and use retraction (Absil et al., 2008) to smoothly map a tangent vector to manifold. For a given point \( U \in S_{n,p} \), let \( G \) be the gradient of \( \mathcal{F}(U) \). A skew symmetric matrix \( A = GU^\top - UG^\top \) is then defined to get the following update in closed form (Wen & Yin, 2013).

\[
\mathbf{V}(\tau) = QU, \quad \text{where } Q = \left( I + \frac{\tau}{2} A \right)^{-1} \left( I - \frac{\tau}{2} A \right)
\]  

Since we seek fast updates on the Stiefel manifold, we resort to this update scheme in designing our metric learning algorithm.

3. Proposed Framework

3.1. Notations

We denote the set of data points \( X = \{ x_1, x_2, \ldots, x_m \} \), with \( x_i \in \mathbb{R}^n, \; i = 1, \ldots, m \) and their corresponding class labels by \( \ell_i \). The \( n \)-dimensional real space is denoted by \( \mathbb{R}^n \) and its positive orthant as \( \mathbb{R}_{+}^n \). As we use pairwise constraints for metric learning, the constraint point pairs are grouped into two sets: \( \mathcal{C}_s \), the set of similar pairs and \( \mathcal{C}_d \), the set of dissimilar pairs. The complete set of constraints is denoted by \( \mathcal{C} = \mathcal{C}_s \cup \mathcal{C}_d \).

3.2. Problem Formulation

Our formulation for metric learning is based on the premise that PSD matrices have nonnegative eigenvalues and orthogonal eigenvectors. Thus, we work with the representation of the rank \( p \) Mahalanobis matrix obtained by its eigendecomposition, \( M_{n \times n} = U \Sigma U^\top \), \( \Sigma = \text{Diag}(\omega) \), where \( U \in S_{n,p} \) is the orthonormal matrix of eigenvectors, and \( \omega \in \mathbb{R}_+^p \) is the vector of eigenvalues. We rewrite the metric learning problem as a joint optimization over \( S_{n,p} \times \mathbb{R}_+^p \) and use \( ||\omega||_2^2 \) as the regularization function, which is equivalent to \( ||M||_F^2 \), the squared Frobenius norm of \( M \).

The convex metric learning problem with the Frobenius norm regularizer is

\[
\min_{M \in S_{n,p}^+} ||M - M_0||_F^2 \tag{2}
\]

subject to \( z_{ij}^\top M z_{ij} \leq s, \; \forall \; i, j \in \mathcal{C}_s \)

\( z_{ij}^\top M z_{ij} \geq d, \; \forall \; i, j \in \mathcal{C}_d \)

where the vectors \( z_{ij} \) are the difference vectors \( x_i - x_j \) obtained from the constraint pairs in \( \mathcal{C}_s \) and \( \mathcal{C}_d \) respectively. \( M_0 \) is the initial Mahalanobis distance matrix, often initialized to identity or the data covariance matrix. Since the problem in (2) could be infeasible, we introduce slack variables \( \xi \) and rewrite the relaxed problem as

\[
\begin{align*}
\min_{w \in \mathbb{R}_+^p, \xi \in \mathbb{R}^{|\mathcal{C}|}} & \quad ||w - w_0||_2^2 + \gamma ||\xi - \xi_0||_2^2 \\
\text{s.t.} & \quad z_{ij}^\top U \text{Diag}(w) U^\top z_{ij} \leq \xi_{ij}, \; \forall \; i, j \in \mathcal{C}_s \\
& \quad z_{ij}^\top U \text{Diag}(w) U^\top z_{ij} \geq \xi_{ij}, \; \forall \; i, j \in \mathcal{C}_d 
\end{align*}
\]  

(3)

where \( w_0 \) is the vector of eigenvalues of \( M_0 \). The initial vector of slack variables \( \xi_0 \) of length \( |\mathcal{C}| \) takes values \( \{ \xi_0 \}_{ij} = \{ s, d \} \) based on whether \( i, j \in \mathcal{C}_s \) or \( i, j \in \mathcal{C}_d \). Note that the problem becomes nonconvex because of the domain of \( U \), which is the Stiefel manifold. The solution to the problem (3) yields \( \hat{w} \) and \( \hat{U} \), which are used to reconstruct the Mahalanobis matrix \( \hat{M} = \hat{U} \text{Diag}(\hat{w}) \hat{U}^\top \).

Intuitively, the solution to (3) gives an orthogonal basis \( \hat{U} \) of the \( p \)-dimensional subspace of \( \mathbb{R}^n \), along with minimal scaling required to satisfy the distance constraints. While we cannot theoretically guarantee good generalization, our experiments in Section 4 show that results are competitive with metric learned by solving (2).

3.3. Algorithm

We solve the problem developed in (3) jointly over \( S_{n,p} \times \mathbb{R}_+^p \) by taking an alternating minimization approach. We initialize the algorithm with the Euclidean metric in a \( p \)-dimensional space with \( w_0 \) as a vector of ones and \( U_0 \) as a randomly picked point on \( S_{n,p} \).

The algorithm alternates between two steps: it solves for \( U \) in (3) while keeping \( w \) fixed and vice versa till the convergence criteria is satisfied.

The optimization problem in (3), with fixed \( U \) is a constrained least square problem. We rewrite the corresponding unconstrained Lagrangian in (4) and obtain updates for \( w \) and \( \xi \) using KKT conditions for single constraint \( (i, j) \in \mathcal{C} \).

\[
\min_{w \in \mathbb{R}_+^p, \xi \in \mathbb{R}^{|\mathcal{C}|}} ||w - w_0||_2^2 + \gamma ||\xi - \xi_0||_2^2 \tag{4}
\]

\[
+ \lambda_{ij}^t y_{ij}(z_{ij}^\top U \text{Diag}(w^t) U^\top z_{ij} - \xi_{ij}^t)
\]

Here \( y_{ij} = -1 \), if \( i, j \in \mathcal{C}_s \) and \( y_{ij} = 1 \) if \( i, j \in \mathcal{C}_d \) and \( \lambda_{ij}^t \geq 0 \) are the Lagrange multipliers. With an updated \( w \), we then solve for \( U \) for the same constraint pair \( (i, j) \). This is achieved by solving the following problem over the Stiefel manifold using updates in (1)

\[
\min_{U \in S_{n,p}} \lambda_{ij}^t y_{ij}(z_{ij}^\top U \text{Diag}(w^t) U^\top z_{ij} - \xi_{ij}^t) \tag{5}
\]

We pick another constraint and repeat the updates (4) and (5) till convergence. Since \( p \) could be as large as
n, and the updates (1) require inversion of a $2p \times 2p$ matrix (Wen & Yin, 2013), we use a block coordinate descent like strategy proposed in (Collins et al., 2014) to speed up this step. The key idea in (Collins et al., 2014) is to parametrize $U$ by a point on a smaller Stiefel manifold. To obtain this parametrization, a set of $k \leq n$ rows $K$, is selected from $U$ to construct a smaller matrix $H_{K,p}$. If $I$ is the set of linearly independent columns of $H$, the parametrization is given as

$$U(V) = \begin{bmatrix} V P^{1/2} & V P^{1/2} R \\ \bar{U}_{k,I} & \bar{U}_{k,\bar{I}} \end{bmatrix}$$

where $P = H_{I,I}^T H_{I,I}$ is positive definite, the $\bar{K}$ and $\bar{I}$ denote the complementary sets of $K$ and $I$ respectively. The matrix $R \in \mathbb{R}^{K \times I}$ is the linear transformation that maps $H_{I,I}$ to $H_{\bar{I},\bar{I}}$ and the orthonormal matrix $V$ is a point on the smaller Stiefel manifold $S_{k,p}$. Collins (2014) show that a descent curve on $S_{k,p}$ gets mapped to the original manifold $S_{n,p}$ by (6) in a direction of descent. As each block of $k$ rows is updated on a smaller Stiefel manifold, we get efficient updates for $U$. Moreover, this block coordinate descent type strategy can be parallelized by using disjoint sets of rows $K_i$ such that $|\cup_i K_i| \leq n$.

4. Experiments

We refer our formulation in (3) as SMML (Stiefel Manifold based Metric learning) and evaluate the learned metric against the Euclidean distance metric in terms of classification accuracy of a 3-nearest neighbor classifier on the UCI benchmark data sets and USPS digits. We compared the run time of SMML to solve (3) with that of SeDuMi solver becomes computationally expensive in terms of memory usage with impractical run times. The accuracy and run time comparisons are summarized in Table 2.

5. Extension

We later, modify the metric learning formulation in (3) and formulate an unconstrained optimization problem. The two-term objective contains a hinge loss function to control distance constraint violations and a $\ell_2$-norm regularizer to ensure smoothness. This formulation optimizes for all the constraints simultaneously instead of solving for only one constraint at every iteration. The formulation is given by

$$\min_{U \in S_{K,p}, w \in \mathbb{R}^p} \sum_{i,j=1}^{m} \left[ y_{ij} (z_{ij}^T U \text{Diag}(w) U^T z_{ij} - b_{ij}) \right]_+ + \alpha \| w - w_0 \|_2^2$$

Here, the term $[x]_+ = \max(0, x)$ is the hinge loss term that captures the degree of violation of constraints, $b_{ij}$ are the corresponding target distances. and $\alpha > 0$ is a

---

### Table 1. Classification Accuracy and Run Time Results

<table>
<thead>
<tr>
<th>Dataset</th>
<th>USPS</th>
<th>Wine</th>
<th>Inosphere</th>
</tr>
</thead>
<tbody>
<tr>
<td># samples</td>
<td>11000</td>
<td>178</td>
<td>351</td>
</tr>
<tr>
<td># constraints ($</td>
<td>I</td>
<td>$)</td>
<td>900</td>
</tr>
<tr>
<td># dimension</td>
<td>256</td>
<td>13</td>
<td>34</td>
</tr>
<tr>
<td># dimension after PCA</td>
<td>114</td>
<td></td>
<td></td>
</tr>
<tr>
<td># Training points</td>
<td>150</td>
<td>45</td>
<td>30</td>
</tr>
<tr>
<td># Testing points</td>
<td>2000</td>
<td>133</td>
<td>148</td>
</tr>
<tr>
<td># classes</td>
<td>10</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>$</td>
<td>K</td>
<td>$</td>
<td>24</td>
</tr>
</tbody>
</table>

**Classification Accuracy %**
- Euclidean: 76.10, 72.3, 69.4
- CVX: 93.1, 94.6, 98
- SMML: 93.1, 96, 97.3

**Run Time (in secs)**
- CVX: 846.3, 7.2, 22.6
- SMML: 346, 13.6, 39.2

### Table 2. Results on USPS digits for different dimensions from PCA

| PCA dimension, $|K|$ | Run time (in mins) / Accuracy(%) |
|----------------------|---------------------------------|
| 38, 10               | CVX: 10/83, 8.7/82               |
| 66, 22               | CVX: 54/80, 17/89.4              |
| 152, 30              | CVX: -30/93.7                    |
Table 3. Comparison of Classification Accuracies of different metric learning approaches

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Our</th>
<th>ITML</th>
<th>LMNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>USPS digits</td>
<td>79.93±1.54</td>
<td>75.27±1.26</td>
<td>76.88±6.98</td>
</tr>
<tr>
<td>MIT scene</td>
<td>60.7±1.54</td>
<td>50.93±1.72</td>
<td>45.15±6.53</td>
</tr>
</tbody>
</table>

Due to space limitations, we are unable to describe the detailed algorithm for optimizing (7). We use an alternating strategy similar to (Liu et al., 2015).

We present our preliminary results on USPS digits and MIT scene datasets in Table 3. We also compare our approach with state of the art metric learning approaches: ITML (Davis et al., 2007) and LMNN (Weinberger & Saul, 2009). For both the datasets we perform dimensionality reduction using PCA while preserving 95% energy.

6. Conclusion

We proposed a metric learning formulation that poses a joint optimization problem over $S_{n,p} \times \mathbb{R}^p$ to find the eigenvectors and eigenvalues of the learned Mahalanobis distance matrix $M$. We took an alternate minimization approach by iteratively updating the eigenvalues and the eigenvectors of $M$ to solve the ensuing nonconvex problem. The proposed method showed competitive performance in classification tasks against convex formulation as well as state of the art methods. Since, our formulation allows the flexibility to replace the regularizer with any convex spectral function, we plan to explore the impact of other functions like log det or Burg entropy in future.

Dual Submission

This paper is primarily a version of our work published in Proceedings of the 1st International Workshop on DIFFerential Geometry in Computer Vision for Analysis of Shapes, Images and Trajectories (DIFF-CV 2015) 3, pages 7.1-7.10. BMVA Press, September 2015. This work also includes our formulation accepted at ICIP 2016 4 and will be available for Open Preview on IEEE Xplore a month before the conference.

References


Liu, Wei, Mu, Cun, Ji, Rongrong, Ma, Shiqian, Smith, John R., and Chang, Shih-Fu. Low-rank similarity metric learning in high dimensions. In *AAAI Conference on Artificial Intelligence (AAAI)*, Austin, Texas, USA, 2015.


