Mach Learn (2013) 92:133–175
DOI 10.1007/s10994-013-5379-y

Geometry preserving multi-task metric learning

Peipei Yang · Kaizhu Huang · Cheng-Lin Liu

Received: 18 November 2012 / Accepted: 2 May 2013 / Published online: 25 May 2013

Abstract In this paper, we consider the multi-task metric learning problem, i.e., the problem of learning multiple metrics from several correlated tasks simultaneously. Despite the importance, there are only a limited number of approaches in this field. While the existing methods often straightforwardly extend existing vector-based methods, we propose to couple multiple related metric learning tasks with the von Neumann divergence. On one hand, the novel regularized approach extends previous methods from the vector regularization to a general matrix regularization framework; on the other hand and more importantly, by exploiting von Neumann divergence as the regularization, the new multi-task metric learning method has the capability to well preserve the data geometry. This leads to more appropriate propagation of side-information among tasks and provides potential for further improving the performance. We propose the concept of geometry preserving probability and show that our framework encourages a higher geometry preserving probability in theory. In addition, our formulation proves to be jointly convex and the global optimal solution can be guaranteed. We have conducted extensive experiments on six data sets (across very different disciplines), and the results verify that our proposed approach can consistently outperform almost all the current methods.

Keywords Multi-task learning · Metric learning · Geometry preserving · von Neumann divergence · Bregman matrix divergence

Editors: Tijl De Bie and Peter Flach.

P. Yang (✉) · K. Huang · C.-L. Liu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences,
95 Zhongguancun East Road, 100190 Beijing, China
e-mail: ppyang84@gmail.com

K. Huang
e-mail: kaser.huang@gmail.com

C.-L. Liu
e-mail: liucl@nlpr.ia.ac.cn

1 Introduction

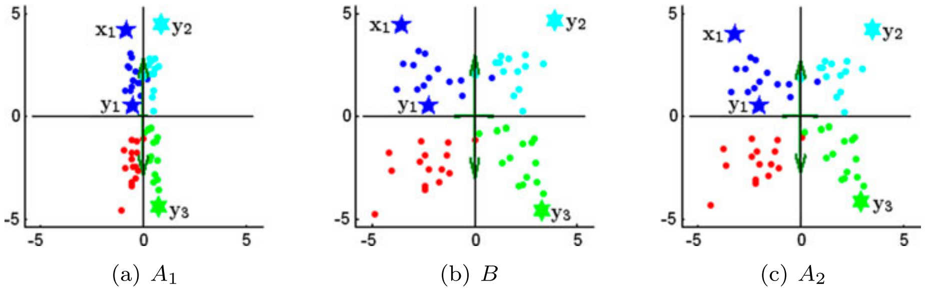
Metric learning has been widely studied in machine learning due to its importance in many machine learning algorithms (Xing et al. 2003; Weinberger and Saul 2009; Davis et al. 2007; Huang et al. 2009, 2011; Ying et al. 2009; Ying and Li 2012). The objective of metric learning is to learn a proper metric function from data, usually a Mahalanobis distance defined as $d_A(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^\top A(\mathbf{x}-\mathbf{y})}$, while satisfying certain extra constraints called side-information, e.g., similar (dissimilar) points should stay closer (further).

In this paper, we consider an extended metric learning problem where there exist several correlated metric learning tasks simultaneously. Two traditional solutions could be exploited for the problem. The first one is to learn a metric for each task individually. Unfortunately, this approach is likely to be over-fitting, especially when the training samples of some tasks are insufficient. On the other hand, the second solution suggests to learn a single global metric for all the tasks. Since the essential discrepancies among the tasks are neglected by this method, the performance is limited. To attack this problem, multi-task learning (MTL) has recently received considerable attention (Caruana 1997; Evgeniou and Pontil 2004; Argyriou et al. 2008; Argyriou and Evgeniou 2008; Zhang et al. 2008; Zhang and Yeung 2010a). MTL learns an individual model for each task but trains them jointly. Joint training of multiple tasks enables information sharing among tasks, which helps improve the performance of each task.

Despite its good performance, MTL has been rarely applied to the multiple metric learning problems. To our best knowledge, only recently Parameswaran and Weinberger (2010), Zhang and Yeung (2010a), and Yang et al. (2011) developed a multi-task metric learning framework separately. Parameswaran and Weinberger (2010) proposed a novel multi-task framework called the *Large Margin Multi-Task Metric Learning (mtLMNN)* which is a combination of the *Large Margin Nearest Neighbor (LMNN)* (Weinberger and Saul 2009) and the *Regularized Multi-task Learning (RegMTL)* (Evgeniou and Pontil 2004). It assumes that the Mahalanobis matrix of each task is composed of a common part and a task-specific part. By minimizing the Frobenius norm of the task-specific part, each metric could be constrained to be similar to a common one so that different tasks may share information from each other. On the other hand, Zhang and Yeung (2010b) proposed to combine the *Multi-task Relationship Learning (MTRL)* (Zhang and Yeung 2010a) with the *Regularized Distance Metric Learning (RDML)* (Jin et al. 2009). By introducing a regularization item with a task covariance matrix, the relationship among tasks can be learned, which provides the potential for better sharing information among the tasks. In addition, Yang et al. (2011) also did some work in this topic by assuming that the metrics of all tasks share a common subspace.

However, all the above mentioned methods have certain limitations. For Yang et al. (2011), since the formulation is not convex, the global optimal solution is not guaranteed. Besides, the assumption of the common subspace may be too strict to be used in some cases. The other two methods exploited vector-based divergence measures to describe the task relationship. Specifically, if we concatenated all columns of each matrix as a vector, in Parameswaran and Weinberger (2010), Frobenius norm between two matrices simply presents the Euclidean distance, while, in Zhang and Yeung (2010a), the regularization applied a matrix-variate normal prior distribution to the vectors. However, we will show that these methods designed for vector variables do not apply to the positive semi-definite Mahalanobis matrices directly and will lead to inaccurate information propagation among tasks. This deficiency will further limit the performance improvement.

For example, the squared Frobenius norm of two Mahalanobis matrices are used to measure the discrepancy of two metrics, but we can show in Fig. 1 that it is not a proper measure



(a) A_1 (b) B (c) A_2

Fig. 1 Illustration of geometry preserving property. The geometry property between d_{A_2} and d_B is better preserved than the one between d_{A_1} and d_B. Besides, the relative distance of $d(\mathbf{x}_1, \mathbf{y}_1)$ and $d(\mathbf{x}_1, \mathbf{y}_2)$ is preserved from B to A_2 but not preserved to A_1

for metrics. There are three figures associated with different distance metrics, determined by a Mahalanobis matrix A_1, B, and A_2 respectively for each figure (from left to right). To visualize the Mahalanobis metric in the Euclidean space (Xing et al. 2003), we transform each point $\mathbf{x}_i$ to $\hat{\mathbf{x}}_i = A^{1/2}\mathbf{x}_i$ when plotting so that the Euclidean distance of any pair of transformed points $d(\hat{\mathbf{x}}_i, \hat{\mathbf{x}}_j)$ is exactly the Mahalanobis distance of the original points $d_A(\mathbf{x}_i, \mathbf{x}_j)$. Geometrically observed, the metric B is obviously more similar to A_2 than to A_1. However, when calculating the similarity using the squared Frobenius norm of the Mahalanobis matrix difference, surprisingly, B is more similar to A_1 than to A_2! This shows that minimizing Frobenius norm of matrix difference cannot preserve the geometry and hence may not be appropriate for measuring the divergence of metrics.

In contrast to the above methods, in this paper, we engage the *Bregman matrix divergence* (Dhillon and Tropp 2008) to design a more general regularized framework for multi-task metric learning. On one hand, this general framework exploits a more general family of matrix divergences. We show that it naturally incorporates the mtLMNN (using the Frobenius norm) as a special case. On the other hand and more importantly, by exploiting a special Bregman divergence called *von Neumann divergence* (Dhillon and Tropp 2008) (denoted by $D_{\mathrm{vN}}(A, B)$) as the regularization, the new multi-task metric learning method has the capability to well preserve the geometry when transferring information from one metric to another. The geometry preserving property is important because (1) data usually live in a geometric vector space in the traditional learning tasks and (2) metric learning is also usually conducted in a geometric vector space, e.g., Euclidean space. In this sense, to guarantee a better performance, it is necessary to preserve the data geometry, e.g., those relative constraints such as sample $\mathbf{x}_i$ should be more similar to sample $\mathbf{x}_j$ than sample $\mathbf{x}_k$, when transferring information among tasks.[1]

We define the *geometry preserving probability* to measure the geometry preserving property of two metrics mathematically from the statistical point of view. Then a series of theoretical analysis is provided to show that our new multi-task metric learning method usually leads to a higher geometry preserving probability and has the capability to better preserve geometry. This enables more appropriate information propagation among tasks and hence provides potential for further raising the performance. In addition to the geometry preserving

[1] We note that there are many applications where either it is not possible to find satisfactory features or they are inefficient for learning purposes in a geometric space. In these cases, it may then be unreasonable to preserve the geometry. However, learning in such domains appears beyond the scope of our paper and hence we leave it as one of the interesting future explorations.

property, the new multi-task framework with the von Neumann divergence remains jointly convex, provided that any convex metric learning is used. This is one of the major advantages against other non-convex formulations, e.g., the model proposed in Yang et al. (2011). Extensive experimental results on one synthetic data set and five real data sets (across very different disciplines) also verify that our proposed algorithm can consistently outperform the current methods. Especially, a toy example in Fig. 4 of Sect. 6.1 can show the advantage of our method more intuitively.

This paper is an extension of our earlier conference paper (Yang et al. 2012), which first proposed the concept of geometry preserving property and used to improve multi-task metric learning problems. This journal version significantly extends the previous paper both theoretically and empirically. It reviews the related methods and summarizes their strengths and weaknesses, explains the motivation in more details, enhances the theoretical analysis in a stricter way with complete proofs of all theorems, and expands the experimental results by comparing with more methods on more datasets.

The rest of this paper is organized as follows. In Sect. 2, we provide the notations used in the paper. In Sect. 3, we review the related work. In Sect. 4, we then present the novel multi-task metric learning framework with Bregman matrix divergence, the concept of geometry preserving probability, the proposed learning method and optimization algorithm. We present theoretical analysis in Sect. 5 and experimental evaluation in Sect. 6. Finally, we give concluding remarks in Sect. 7.

2 Notations and problem definition

In this section, we first present the notations used in the paper and then introduce the problem definition of multi-task metric learning.

A *metric* defined on set $\mathbb{X}$ is a *function* $d : \mathbb{X} \times \mathbb{X} \to \mathbb{R}_+ \triangleq [0, +\infty)$ satisfying positiveness, symmetry, and triangle inequality (Burago et al. 2001). In this paper, we focus on the Mahalanobis metric defined on $\mathbb{R}^m$ by a *symmetric positive semi-definite (SPSD)* matrix A as $d_A(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^\top A(\mathbf{x} - \mathbf{y})}$ where A is called Mahalanobis matrix. Denoting the set composed of all metrics on $\mathbb{X}$ by $\mathcal{F}_\mathbb{X}$ and given any pair of metrics $d_A, d_B \in \mathcal{F}_\mathbb{X}$, a divergence function $D : \mathcal{F}_\mathbb{X} \times \mathcal{F}_\mathbb{X} \to \mathbb{R}_+$ is defined to measure the discrepancy of d_A and d_B. Since the Mahalanobis metric d_A is parameterized by the Mahalanobis matrix A, we denote $D(d_A, d_B) \triangleq D(A, B)$ for short.

Assume that there are T related metric learning tasks. For each task-t, its training data set $\mathcal{X}_t = \{\mathbf{x}_{tk} \in \mathbb{R}^m\}_{k=1}^{N_t}$ contains N_t data points where m is the dimension. The side-information defining a set of constraints on the learned metric d_t can be generally formulated as $d_t \in \mathcal{C}_t(\mathcal{X}_t)$. For instance, in the *Generalized Sparse Metric Learning (GSML)* (Huang et al. 2009) and the LMNN, $\mathcal{C}_t$ is defined as a triplet set $\mathcal{T}_t = \{(i, j, k)\}$ which provides side-information as relative constraints such that $\mathbf{x}_{ti}$ is more similar to $\mathbf{x}_{tj}$ than to $\mathbf{x}_{tk}$ under the new metric and thus

$$\mathcal{C}_t(\mathcal{X}_t) = \big\{d \in \mathcal{F}_\mathbb{X} \mid d(\mathbf{x}_{ti}, \mathbf{x}_{tj}) \le d(\mathbf{x}_{ti}, \mathbf{x}_{tk}),\ \forall (i, j, k) \in \mathcal{T}_t\big\}.$$

In the *Informative Theoretical Metric Learning (ITML)* (Davis et al. 2007), $\mathcal{C}_t$ is defined as a similar pair set $\mathbb{S}_t$ and a dissimilar pair set $\mathbb{D}_t$ which provides side-information such that similar (dissimilar) pairs should stay closer (further) than a upper bound u (lower bound l) respectively and thus

$$\mathcal{C}_t(\mathcal{X}_t) = \left\{ d \in \mathcal{F}_\mathbb{X} \,\middle|\, \begin{aligned} & d(\mathbf{x}_{ti}, \mathbf{x}_{tj}) \le u,\ \forall (i, j) \in \mathbb{S}_t; \\ & d(\mathbf{x}_{ti}, \mathbf{x}_{tj}) \ge l,\ \forall (i, j) \in \mathbb{D}_t. \end{aligned} \right\}.$$

The objective of multi-task metric learning is to learn T proper Mahalanobis matrices $\{A_t\}_{t=1}^T$ jointly, which is significantly different from single-task metric learning where each Mahalanobis matrix is learned independently.

3 Related work

There have been some attempts to combine multi-task learning with metric learning. Based on different assumptions about the relationship among tasks, the researchers proposed some interesting models of multi-task metric learning.

3.1 Multi-task large margin metric learning

The first multi-task metric learning method is the mtLMNN model proposed by Parameswaran and Weinberger (2010). Motivated by the RegMTL (Evgeniou and Pontil 2004), the mtLMNN assumes that the Mahalanobis matrix of the t-th task is composed of a common part and a task-specific part as $A_t = A_0 + \hat{A}_t$. Exploiting further the squared Frobenius norm (Horn and Johnson 1985) of the task-specific part $\|\hat{A}_t\|_{\mathrm{F}}^2$ as the regularization term, mtLMNN encourages the similarity between each task and a common one. This approach indeed shows better performance in several real data sets. However, this method suffers from two drawbacks which we explain at the end of Sect. 4.1, which will further limit its performance in real applications.

3.2 Zhang and Yeung's method

Zhang and Yeung (2010b) proposed a multi-task metric learning approach by assuming the matrix composed of vectorized Mahalanobis matrices of all tasks follows a matrix-variate normal distribution (Zhang and Yeung 2010a; Gupta and Nagar 2000). It first concatenates all columns of each A_t to form a vector $\tilde{A}_t = \mathrm{vec}(A_t)$ and then engages the MTRL (Zhang and Yeung 2010a) regularization $\tilde{A}\Omega^{-1}\tilde{A}^\top$ to couple different tasks, where $\tilde{A} = [\mathrm{vec}(A_1), \dots, \mathrm{vec}(A_T)]$. It applies a matrix-variate normal prior distribution

$$q(\tilde{A}) = \mathcal{MN}_{m^2\times T}(\tilde{A}|\mathbf{0}_{m^2\times T}, \mathbf{I}_{m^2} \otimes \Omega)$$

to $\tilde{A}_t$'s (Zhang and Yeung 2010a) and the task relationship Ω can finally be obtained together with all the metrics. This approach has demonstrated some desirable properties as the task relationship can be learned automatically, but there are two irrationalities of the prior distribution applied to $\tilde{A}$:

- The expectation of each $\tilde{A}_t$ is a zero vector, which is apparently designed for vector-based variables rather than Mahalanobis matrices being symmetric semi-positive definite. For example, A and $-A$ are assigned with equal prior probability, which is improper since at most one of them is possible to be a feasible Mahalanobis matrix.
- Vectorization of a matrix discards some structure information.

Moreover, the authors surprisingly failed to validate it empirically.

Actually, mtLMNN also applies a multi-variate normal distribution to the vectorized Mahalanobis matrices. In contrast to Zhang and Yeung (2010b)'s method which predefines the mean and learns the task relationship, mtLMNN predefines the task relationship Ω as the Laplacian matrix of an all connected graph (Chung 1997).

3.3 Multi-task metric learning based on common subspace

Yang et al. (2011) proposed their multi-task metric learning method based on the assumption that all the metrics share a common low-dimensional subspace. Supposing $A_t = L_t^\top L_t$ and the transformation matrix L_t has the decomposition $L_t = R_t L_0$, all tasks are coupled by the common matrix L_0, which has fewer rows than its columns. Hence it actually defines a common subspace for all the tasks, while R_t defines the metric in this subspace for each task. With alternating optimization, the subspace L_0 and all the metrics R_t can be solved simultaneously. However, this assumption is sometimes too strict. In addition, this model involves a non-convex optimization and hence cannot guarantee the global solution.

4 Geometry preserving multi-task metric learning

In this section, we first detail our proposed novel framework, and then show the importance of preserving geometry among samples when sharing the side-information among tasks. The concept of *geometry preserving probability* is then proposed to provide a mathematical criterion that measures the capability to preserve the geometry relationship, i.e., the relative distance of samples between two metrics. Following that, we introduce our method that exploits von Neumann divergence to regularize the relationship among multiple tasks. Finally, we present a practical algorithm to solve the involved optimization problem.

4.1 General framework

In this section, we propose a general framework for multi-task metric learning including mtLMNN as a special case.

Assume that a common metric d_c is defined and the metric of each (the t-th) task d_t is correlated with d_c by a regularization $D(d_t, d_c)$. All the metrics are coupled by this common metric. In the case of the Mahalanobis metric, the regularization can be also written as $D(A_t, B)$, where A_t and B correspond to the t-th task and the common one respectively. Then the novel framework can be formulated as

$$\begin{aligned} \min_{\{A_t\},B} \quad & \sum_t \big(L(A_t, \mathcal{X}_t) + \gamma D(A_t, B) \big) + \gamma_0 D(A_0, B) \\ \text{s.t.} \quad & A_t \in \mathcal{C}_t(\mathcal{X}_t), \\ & A_t \succeq \mathbf{0}, \end{aligned} \tag{1}$$

where L is a loss function of the training samples depending on side-information and the metric learning method, $\mathcal{X}_t$ represents the set of training samples of the t-th task, $D(\cdot,\cdot)$ is the divergence function to correlate two metrics, and $\mathcal{C}_t(\mathcal{X}_t)$ is the set of feasible A_t determined by side-information. The predefined metric A_0 provides a prior for the common metric and we can usually use the Euclidean distance, i.e., $A_0 = \mathbf{I}_m$. In a lot of cases, there may not exist a feasible solution to strictly satisfy all the constraints defined as $\mathcal{C}_t(\mathcal{X}_t)$ and thus the *soft constraints* are used instead by reformulating the inequality constraints as loss functions. For example, the constraint $d_A^2(\mathbf{x}_i, \mathbf{x}_k) - d_A^2(\mathbf{x}_i, \mathbf{x}_j) \geq 1$ is reformulated as a loss $[1 + d_A^2(\mathbf{x}_i, \mathbf{x}_j) - d_A^2(\mathbf{x}_i, \mathbf{x}_k)]_+$ where $[z]_+ = \max(z, 0)$. Then, denoting the loss function including the soft constraints as $\tilde{L}(A_t, \mathcal{X}_t)$, the framework becomes

$$\min_{\{A_t\},B} \sum_t \big(\tilde{L}(A_t, \mathcal{X}_t) + \gamma D(A_t, B) \big) + \gamma_0 D(A_0, B) \quad \text{s.t. } A_t \succeq \mathbf{0}.$$

In our framework, all metrics are correlated with each other because the model assumes that each d_{A_t} is encouraged to be similar to a common metric d_B by minimizing $D(A_t, B)$. Thus it plays a role of measuring the discrepancy of two metrics so that the less $D(A_t, B)$ is, the more closely A_t and B are correlated. Therefore, by minimizing $D(A_t, B)$, the information is enforced to be shared between A_t and B, and the definition of $D(\cdot, \cdot)$ determines the type of shared information. Since multi-task learning improves the performance of each task by utilizing the information propagated from others, the choice of $D(\cdot, \cdot)$ is critical to the performance of this framework.

There is a family of discrepancy measures for two Hermitian matrices called *Bregman matrix divergence* (Dhillon and Tropp 2008), which is defined as

$$D_\phi(A, B) = \phi(A) - \phi(B) - \operatorname{tr}\big(\big(\nabla\phi(B)\big)^\top (A - B)\big), \tag{2}$$

where $\phi : \mathcal{H} \to \mathbb{R}$ is a strictly convex, differentiable *generating function* of a Hermitian matrix variable, and $\operatorname{tr}(A)$ is the *trace* of A. Furthermore, if $\phi(A)$ depends only on the eigenvalues of A, it is called a *spectral function* (Lewis 1996), and $D_\phi(A, B)$ is the *spectral Bregman matrix divergence* (Kulis et al. 2009). In this case, ϕ can be written as a composition $\phi(A) = (\varphi \circ \lambda)(A)$, where $\lambda(A)$ is the function that lists the eigenvalues in algebraically decreasing order and φ is a strictly convex function on $\mathbb{R}^m$.

By choosing different φ, we obtain some famous types of matrix divergences (Kulis et al. 2009). If the squared 2-norm $\varphi(\mathbf{x}) = \mathbf{x}^\top\mathbf{x} = \|\mathbf{x}\|^2$ is used, we have $\phi(A) = \|A\|_{\mathrm{F}}^2$ and $D_\phi(A, B) = \|A - B\|_{\mathrm{F}}^2$ is the squared *Frobenius norm* of their difference; if the entropy $\varphi(\mathbf{x}) = \sum_i x_i \log x_i - x_i$ is used, we have $\phi(A) = \operatorname{tr}(A \log A - A)$[2] and $D_\phi(A, B)$ is the *von Neumann divergence*, which we will discuss in detail later.

Note that the mtLMNN is a special case of our framework with $D_\phi(A_t, B) = \|A_t - B\|_{\mathrm{F}}^2$ and replacing $A_t \succeq \mathbf{0}$ with $A_t \succeq B \succeq \mathbf{0}$. By rewriting it in this form, its main drawbacks are much clearer:

1. The constraints $A_t \succeq B$ are unnecessarily strong for A_t to be a Mahalanobis matrix, which implies that the distance of any task has to be further than the distance defined by the common metric.
2. Frobenius norm of Mahalanobis matrix difference is inadequate to measure the discrepancy of two metrics, and thus minimizing the Frobenius norm of matrix difference cannot preserve the geometry relation of data defined by the metrics. We have illustrated it with an example and will explain it theoretically in Sect. 5.

4.2 Regularization and geometry preserving probability

In this section, we show our motivation to define a proper regularization $D(A_t, B)$ which enables side-information propagate among tasks more appropriately. Then the concept of geometry preserving probability is proposed to measure whether the side-information is well propagated.

On one hand, in this framework, a smaller $D(A_t, B)$ implies more side-information shared. Noting that the metric is learned by satisfying a set of constraints from side-information, metric learning can also be regarded as a process to embed the side-information into the learned metric. Thus closely correlated metrics should contain similar side-information and minimizing $D(A_t, B)$ should encourage side-information to propagate between A_t and B.

[2] The $\log A$ denotes the *matrix logarithm* whose definition is given in Sect. 4.3.

On the other hand, the side-information is usually formulated as a set of constraints on the relative distance of the samples (Ying and Li 2012). For example, the GSML and the LMNN define the side-information directly by constraints on the relative distances of samples in a triplet set. For the ITML, although it defines an upper bound for the distances of similar pairs and a lower bound for the distances of dissimilar pairs respectively, a set of constraints on their relative distances are also implicitly defined by the relation of the two bounds.

From the above observations, a proper $D(\cdot,\cdot)$ for multi-task metric learning should have the following property: the less $D(A_t, B)$ is, the more constraints about relative distances are satisfied by both A_t and B. Focusing on the t-th task and fixing the common metric B, we obtain the subproblem

$$\min_{A_t} \tilde{L}(A_t, \mathcal{X}_t) + \gamma D(A_t, B) \quad \text{s.t. } A_t \succeq 0, \tag{3}$$

which aims to find an A_t that is correlated with B while satisfying the side-information of its own task. According to the above discussion, it is equivalent to solving such a metric A_t: on one side, it satisfies the constraints from side-information of the t-th task; on the other side, it preserves the geometry relationship (relative distances) of the samples measured by B as better as possible, which we call as "*geometry preserving property*".

To illustrate the geometry preserving property, recall the example shown in Fig. 1. There are two pairs of randomly selected points $(\mathbf{x}_1, \mathbf{y}_1)$, $(\mathbf{x}_1, \mathbf{y}_2)$. Since $d_B(\mathbf{x}_1, \mathbf{y}_1) < d_B(\mathbf{x}_1, \mathbf{y}_2)$, if we want a metric d_A which is similar to d_B, it is desirable that d_A makes the same judgement on the relative distance, i.e. $d_A(\mathbf{x}_1, \mathbf{y}_1) < d_A(\mathbf{x}_1, \mathbf{y}_2)$. Obviously, such a relative distance relationship for $(\mathbf{x}_1, \mathbf{y}_1)$, $(\mathbf{x}_1, \mathbf{y}_2)$ is preserved between A_2 and B but not preserved between A_1 and B. Analogously, there are also two pairs $(\mathbf{x}_1, \mathbf{y}_1)$, $(\mathbf{x}_1, \mathbf{y}_3)$, whose relative distance relationship is preserved between both A_1, B and A_2, B. Since there are more relationships preserved for A_2, B, we say the geometry preserving property of them is better, which is also consistent with our intuition that B is more similar to A_2 than to A_1.

Based on the idea, we propose the concept of *geometry preserving probability* to measure the geometry preserving property mathematically. It is defined as the probability that the relative distance of arbitrary two pairs of random points is preserved for the two metrics.

Definition 1 (Geometry Preserving Probability) Suppose $\mathbf{x}_1, \mathbf{y}_1 \in \mathbb{X}$ and $\mathbf{x}_2, \mathbf{y}_2 \in \mathbb{X}$ are two pairs of random points following a certain distribution defined by probability density $f(\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2)$.

If two metrics d_A and d_B defined on $\mathbb{X}$ are used to compare the distances between each pair of points $d(\mathbf{x}_1, \mathbf{y}_1)$ and $d(\mathbf{x}_2, \mathbf{y}_2)$, the probability of that d_A and d_B make the same judgement about their relative distance is called *geometry preserving probability* of metrics d_A and d_B with distribution f. It is denoted by $\mathrm{PG}_f(d_A, d_B)$ (**P**robability of **G**eometry Preserving) with mathematical formulation shown in (4).

$$\begin{aligned}\mathrm{PG}_f(d_A, d_B) = {} & \mathrm{P}\big[d_A(\mathbf{x}_1, \mathbf{y}_1) > d_A(\mathbf{x}_2, \mathbf{y}_2) \ \wedge \ d_B(\mathbf{x}_1, \mathbf{y}_1) > d_B(\mathbf{x}_2, \mathbf{y}_2)\big] \\ & + \mathrm{P}\big[d_A(\mathbf{x}_1, \mathbf{y}_1) < d_A(\mathbf{x}_2, \mathbf{y}_2) \ \wedge \ d_B(\mathbf{x}_1, \mathbf{y}_1) < d_B(\mathbf{x}_2, \mathbf{y}_2)\big] \\ & + \mathrm{P}\big[d_A(\mathbf{x}_1, \mathbf{y}_1) = d_A(\mathbf{x}_2, \mathbf{y}_2) \ \wedge \ d_B(\mathbf{x}_1, \mathbf{y}_1) = d_B(\mathbf{x}_2, \mathbf{y}_2)\big],\end{aligned} \tag{4}$$

where $(\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2) \sim f$ and $\wedge$ denotes the logical "*and*" operator.

By this definition, the higher $\mathrm{PG}_f(d_A, d_B)$ is, the better the geometry relation is preserved between d_B and d_A. In the example of Fig. 1, using randomly generated samples, the PG can be estimated[3] as $\mathrm{PG}_f(d_{A_1}, d_B) \approx 0.805 < \mathrm{PG}_f(d_{A_2}, d_B) \approx 1.000$ for some distribution f, which shows the geometry is better preserved between B and A_2 than between B and A_1.

In the following parts, we will discuss the proposed method with von Neumann divergence. Theoretical analysis is given in Sect. 5, which will show that our method is more likely to make $\mathrm{PG}_f(d_A, d_B)$ higher and thus can better preserve geometry.

4.3 Multi-task metric learning with von Neumann divergence

We propose to use the *von Neumann divergence* (Dhillon and Tropp 2008; Kulis et al. 2009) as the regularization in framework (1) and obtain our multi-task metric learning method.

Assuming the spectral decomposition of A is $A = V\Lambda V^\top$, the *matrix logarithm* of A is defined as $\log A = V \log \Lambda V^\top = \sum_i \log\lambda_i(\mathbf{v}_i\mathbf{v}_i^\top)$, where $\log \Lambda$ is the diagonal matrix containing the logarithm of eigenvalues.

Then the *von Neumann divergence* is defined as

$$D_{\mathrm{vN}}(A, B) = \mathrm{tr}(A \log A - A \log B - A + B). \tag{5}$$

If either A or B is low-rank, the von Neumann divergence is unavailable due to its zero eigenvalues. In this case, the von Neumann divergence is defined as

$$D_{\mathrm{vN}}(A, B) = D_{\mathrm{vN},U}(A, B) = D_{\mathrm{vN}}\big(U^\top AU, U^\top BU\big),$$

where U is an $m \times r$ column orthogonal matrix such that $\mathrm{range}(B) \subseteq \mathrm{range}(U)$, and this definition is independent of the choice of U (Kulis et al. 2009). Please refer to Kulis et al. (2009) for more detail about the treatment of the low-rank case.

The von Neumann divergence has been widely used in quantum information theory (Nielsen and Chuang 2010). It plays the role of relative entropy between two quantum density operators, which are mathematically represented as SPSD matrices just like the Mahalanobis matrices. Exploiting the von Neumann divergence as the regularization between Mahalanobis matrices A and B, the geometry relationship of samples measured by B is more liable to be preserved as measured by A. More strictly, it will encourage a higher geometry preserving probability $\mathrm{PG}_f(A, B)$. We will detail the theoretical analysis in Sect. 5.

The von Neumann divergence has a nice property that it is jointly convex with both two arguments (Tropp 2012; Bauschke and Borwein 2001) as shown in Theorem 1.

Theorem 1 (Joint convexity of von Neumann divergence) *The von Neumann divergence* (5) *is* jointly convex, *which means that for SPD matrices* $\{A_i, B_i\}_{i=1}^n$ *and* $\{p_i \in [0, 1]\}_{i=1}^n$ *satisfying* $\sum_i p_i = 1$, *the following inequality holds.*

$$D_{\mathrm{vN}}\left(\sum_i p_i A_i, \sum_i p_i B_i\right) \le \sum_i p_i D_{\mathrm{vN}}(A_i, B_i).$$

This theorem can be derived from the *Lindblad's Theorem* (Lindblad 1973). A detailed proof can be seen in Tropp (2012), Bauschke and Borwein (2001).

Therefore, given a convex metric learning algorithm, it can be extended to a jointly convex multi-task metric learning problem using our method. We solve it by the alternating optimization method. At the initial state, B is set to A_0. Then each A_t and B are solved alternately with other variables fixed. A global optimal solution is guaranteed due to its convex nature. We elaborate the optimization in the next subsection.

[3] Please refer to Appendix A.4 for the detail of the procedure to estimate the PG.

4.4 Optimization

4.4.1 Fix B and optimize on A_t's

When B is fixed, the optimization problem about A_t's is decoupled into T individual single-task metric learning subproblems (3) and there is an additional regularization $D_{\mathrm{vN}}(A_t, B)$ for each of them. Given that the original metric learning optimization is convex, this subproblem is also convex and they can be solved separately.

If the problem is solved utilizing a gradient descent method or subgradient method, the gradient $\frac{\partial D_{\mathrm{vN}}}{\partial A_t} = \log A_t - \log B$ is needed in each step. In this paper, we apply our multi-task framework to the LMNN (Weinberger and Saul 2009) metric learning algorithm which proved effective in many applications. In each gradient descent step of our algorithm, the additional calculation is just the matrix logarithm of A_t's and B where a spectral decomposition is needed. However, in order to project the obtained solution into the SPSD cone, the LMNN algorithm itself includes the spectral decomposition in each updating step. Thus the calculation of the matrix logarithm can use this result directly. Then the additional calculation is only the logarithm of the eigenvalues and a matrix multiplication.

It should be again carefully treated when A_t is low-rank, which means the current solution moves to the boundary the of domain of $D_{\mathrm{vN}}(A, B)$. The gradient cannot be calculated directly on these points and we can resort to the subspace spanned by the eigenvectors corresponding to the positive eigenvalues. Please refer to Sect. 4 of Kulis et al. (2009) for more details.

4.4.2 Fix A_t and optimize on B

The optimization problem about B with all A_t's fixed is

$$\min_B \sum_t \gamma D_{\mathrm{vN}}(A_t, B) + \gamma_0 D_{\mathrm{vN}}(A_0, B) \quad \text{s.t. } B \succeq 0.$$

This problem is just a special case of Proposition 1 in Banerjee et al. (2005) where the optimal solution is called *Bregman representative*, but in the case of matrix variables. Here we generalize this result into the case of symmetric matrices where the optimal solution is also the weighted average of A_t's as shown in Theorem 2.

Theorem 2 (Bregman matrix representative) *Let $\{X_i\}_{i=1}^n$ be a set of symmetric matrices and $\{p_i\}_{i=1}^n$ form a probability distribution where $\sum_i p_i = 1$. Then for any Bregman divergence, the problem*

$$\min_Y \sum_i p_i D_\phi(X_i, Y)$$

has a unique minimizer given by $Y^* = \sum_i p_i X_i$

Proof The function to be minimized is $J_\phi(Y) = \sum_i p_i D_\phi(X_i, Y)$. Let $\bar{X} = \sum_i p_i X_i$, then for $\forall Y$,

$$\begin{aligned} & J_\phi(Y) - J_\phi(\bar{X}) \\ & \quad = \sum_i p_i D_\phi(X_i, Y) - \sum_i p_i D_\phi(X_i, \bar{X}) \end{aligned}$$

$$= \phi(\bar{X}) - \phi(Y) - \mathrm{tr}\Big(\big(\nabla\phi(Y)\big)^\top\Big(\sum_i p_i X_i - Y\Big)\Big) + \mathrm{tr}\Big(\big(\nabla\phi(\bar{X})\big)^\top\Big(\sum_i p_i X_i - \bar{X}\Big)\Big)$$

$$= \phi(\bar{X}) - \phi(Y) - \mathrm{tr}\big(\big(\nabla\phi(Y)\big)^\top(\bar{X} - Y)\big)$$

$$= D_\phi(\bar{X}, Y) \ge 0.$$

Since ϕ is strictly convex, the equality holds only when $Y^* = \bar{X} = \sum_i p_i X_i$. □

It is very interesting that this result *does not depend* on the choice of ϕ. Then, in our problem, the optimal solution of the common metric is

$$B = \frac{\gamma \sum_t A_t + \gamma_0 A_0}{\gamma T + \gamma_0}.$$

When von Neumann divergence is used, since $\forall A_t \succeq 0$ and $\gamma, \gamma_0 > 0$, the constraint $B \succeq 0$ is automatically satisfied.

4.4.3 Convergence of the alternating optimization

Our alternating optimization approach is indeed a block coordinate descent method (Tseng 1988, 2001; Friedman et al. 2007). In this section, we show that this method will converge to the optimal solution by alternating optimization if von Neumann divergence or squared Frobenius norm is used and the prior is chosen as $A_0 = \mathbf{I}_m$.

Tseng (2001) did an in-depth research about the block coordinate descent method and presented a condition to guarantee the convergence of this algorithm. The objective function to be optimized in this paper has the following special form:[4]

$$f(\mathbf{x}_1, \dots, \mathbf{x}_N) = f_0(\mathbf{x}_1, \dots, \mathbf{x}_N) + \sum_{k=1}^{N} f_k(\mathbf{x}_k)$$

for some $f_0 : \mathbb{R}^{n_1+\cdots+n_N} \to \mathbb{R} \cup \{\infty\}$ and some $f_k : \mathbb{R}^{n_k} \to \mathbb{R} \cup \{\infty\}$.

The condition to guarantee convergence of the coordinate descent method is proposed in Proposition 5.1 of Tseng (2001) with a series of assumptions:

(B1) f_0 is continuous on $\mathrm{dom} f_0$.
(B2) For each $k \in \{1, \dots, N\}$ and $(\mathbf{x}_j)_{j \ne k}$, the function $\mathbf{x}_k \mapsto f(\mathbf{x}_1, \dots, \mathbf{x}_N)$ is quasiconvex and hemivariate (Tseng 2001).
(B3) $f_0, f_1, \dots, f_N$ are lsc (lower semicontinuous).

(C1) $\mathrm{dom} f_0$ is open and f_0 tends to ∞ at every boundary point of $\mathrm{dom} f_0$.
(C2) $\mathrm{dom} f_0 = Y_1 \times \cdots \times Y_N$, for some $Y_k \subseteq \mathbb{R}^{n_k}$, $k = 1, \dots, N$.

There are $N = T + 1$ coordinate blocks in our problem as $\mathbf{x}_i = A_i$, $i = 1, \dots, T$ and $\mathbf{x}_{T+1} = B$. The function with respect to all the variables is $f_0(\mathbf{x}_1, \dots, \mathbf{x}_{T+1}) = \gamma \sum_{i=1}^{T} D_\phi(\mathbf{x}_i, \mathbf{x}_{T+1})$ and the separable functions are $f_i(\mathbf{x}_i) = \tilde{L}(\mathbf{x}_i, \mathcal{X}_i) + \delta_{X \succeq 0}(A_i)$, $i = 1, \dots, T$ and $f_{T+1}(\mathbf{x}_{T+1}) = \gamma_0 D_\phi(A_0, \mathbf{x}_{T+1})$, where $\delta_{X\succeq 0}(A)$ is the *characteristic function* (Wikipedia 2012) of the positive semi-definite cone: $\delta_{X\succeq 0}(A) = 0$ if $A \succeq 0$ and $\delta_{X\succeq 0}(A) = +\infty$ otherwise.

[4] We use the same notations as Tseng (2001) in this subsection, which may cause some confusions with other sections of this paper.

Then we can check the conditions above:

(B1) Both $\|A-B\|_F^2$ and $D_{\text{vN}}(A,B)$ are continuous and thus (B1) is satisfied.

(B2) $\tilde{L}(\mathbf{x}_i,\mathcal{X}_i)$ is convex because a convex metric learning algorithm is used, and $\delta_{X\succeq 0}(A)$ is also convex due to the convexity of the positive semi-definite cone (Boyd and Vandenberghe 2004). Thus $f_i(\mathbf{x}_i)$ is convex for $i=1,\dots,T$. On the other hand, f_0 is convex due to the strict convexity of $\|A-B\|_F^2$ and $D_{\text{vN}}(A,B)$. Then it is straightforward to obtain that f is quasiconvex. It is also not difficult to check that f is hemivariate (Tseng 2001) and thus (B2) is satisfied.

(B3) We choose the metric learning algorithm with continuous objective functions $\tilde{L}$ in this paper and both $\|A-B\|_F^2$ and $D_{\text{vN}}(A,B)$ are continuous. Because $\{X|X\succeq 0\}$ is a closed set, the indicator function $\delta_{X\succeq 0}$ is lsc (Wikipedia 2013b). Thus $f_0,f_1,\dots,f_N$ are all lsc and (B3) is satisfied.

(C2) If $\|A-B\|_F^2$ is used, the domain of each coordinate block is $\mathbb{R}^{m\times m}$, and $\text{dom} f_0 = Y_1\times\cdots\times Y_N$ where $Y_i=\mathbb{R}^{m\times m}, \forall i=1,\dots,T+1$. If $D_{\text{vN}}(A,B)$ is used, each variable should satisfy $A_i\succeq 0$ and $\mathbf{C}(A_i)\subseteq\mathbf{C}(B)$ where $\mathbf{C}(X)$ is the column space of X (Kulis et al. 2009). This seems to make the domains of coordinate blocks dependent with each other. However, since we choose A_0 as the identity matrix in our algorithm, it guarantees B to be a full-rank matrix and thus $\mathbf{C}(A_i)\subseteq\mathbf{C}(B)$ always holds for any i. Then the dependency of variables is decoupled and $\text{dom} f_0 = Y_1\times\cdots\times Y_N$ where $Y_i=\{X\in\mathbb{R}^{m\times m}|X\succeq 0\}, \forall i=1,\dots,T+1$. This proves that (C) is satisfied.

We have shown in the above that $f, f_0, f_1,\dots,f_N$ satisfy Assumptions B1–B3 and f_0 satisfies Assumption C2. In our alternating optimization algorithm, the *cyclic rule* is used which is a special case of the *essentially cyclic rule* (Tseng 2001). Moreover, both the loss $\tilde{L}$ and the Bregman matrix divergence are always non-negative and thus lower bound exists. Then by Proposition 5.1 of Tseng (2001), the algorithm is guaranteed to converge to a minimum point of f.

5 Theoretical analysis of geometry preserving property

In this section, we present a series of theoretical analysis to justify our proposed multi-task metric learning approach has the capability to better preserve data geometry. Before the analysis, we define the concepts of *scale vector* which characterizes the scale property of a metric, and *scale extractor* which is an operator transforming a metric to a *scale vector*. This provides a tool to analyze the relationship between the geometry preserving probability and the Bregman matrix divergence.

In general, the relationship between the geometry preserving probability and Bregman matrix divergence is established in three steps.

1. $\text{PG}_f(d_A,d_B)$ and $\mathcal{E}(A,B)$ are linked: a higher geometry preserving probability $\text{PG}_f(d_A,d_B)$ usually accompanies with a smaller $\mathcal{E}(A,B)$ which is an integration defined with scale vectors in all directions.
2. $D_\phi(A,B)$ and $\mathcal{D}_\varphi(A,B)$ are linked: the Bregman matrix divergence $D_\phi(A,B)$ provides an upper bound for the corresponding Bregman divergence of scales $D_\varphi(\rho_W^A,\rho_W^B)$. Therefore, minimizing $D_\phi(A,B)$ has the effect to minimize $\mathcal{D}_\varphi(A,B)$ which is an integration of Bregman divergence of scales.
3. $\text{PG}_f(d_A,d_B)$ and $D_{\text{vN}}(A,B)$ are linked by $\mathcal{E}(A,B)$ and $\mathcal{D}_{\text{KL}}(A,B)$: when the difference of ρ_W^A and ρ_W^B is small, which is usually satisfied in multi-task problems, $\mathcal{E}(A,B)$

and $\mathcal{D}_{\text{vN}}(A,B)$ are more consistent, which means a smaller (greater) $\mathcal{E}(A,B)$ usually accompanies with a smaller (greater) $\mathcal{D}_{\text{KL}}(A,B)$. Therefore, by minimizing $D_{\text{vN}}(A,B)$, the $\mathcal{D}_{\text{KL}}(A,B)$ is minimized, which furthermore leads to a smaller $\mathcal{E}(A,B)$ implying a higher $\text{PG}_f(d_A,d_B)$ ultimately.

5.1 Scale vector and scale extractor

The concept of *scale* is used to capture the scale (amplified or squashed) property or give an approximate representation of a metric. It translates a metric defined on the complicated functional space $\mathcal{F}_{\mathbb{X}}$ into a simple real vector which contains the most important information of the metric.

Our motivation comes from the following fact. The essential role of a metric is to define the distance for any pair of points in the space. Given any pair of points $\forall \mathbf{x},\mathbf{y}\in\mathbb{X}$, if two *metrics* d_A and d_B are similar, the distances they give, i.e. $d_A(\mathbf{x},\mathbf{y})$ and $d_B(\mathbf{x},\mathbf{y})$, are expected to be similar. This motivates us to measure the similarity between two metrics d_A, d_B in such a way:

1. Select a set of pairs of points $\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$ and measure their distances with the two metrics respectively $\{d_A(\mathbf{x}_i,\mathbf{y}_i), d_B(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^n$;
2. For each metric d_M ($M=A$ or B), use a vector $\rho^M=[f(d_M(\mathbf{x}_1,\mathbf{y}_1))\ \dots\ f(d_M(\mathbf{x}_n,\mathbf{y}_n))]$ as its representation;
3. Since ρ^A and ρ^B are both vectors, it's much easier to define the similarity of them, which can then be used to estimate the similarity of d_A and d_B.

If the metric d is *translation invariant*, i.e., $d(\mathbf{x}+\mathbf{w},\mathbf{y}+\mathbf{w})=d(\mathbf{x},\mathbf{y})$, $\forall \mathbf{x},\mathbf{y},\mathbf{w}\in\mathbb{X}$, such as Mahalanobis metric, we can always translate $\mathbf{x}_i$ to the origin and briefly denote $d(\mathbf{x}_i,\mathbf{y}_i)=d(\mathbf{z}_i,0)\triangleq d(\mathbf{z}_i)$ where $\mathbf{z}_i=\mathbf{x}_i-\mathbf{y}_i$. Then we can imagine that d defines a ruler in each direction $\mathbf{z}_i$ and the most important properties are "the scales of these rulers". Based on this idea, we propose the concept of *scale* as a representation of a metric.

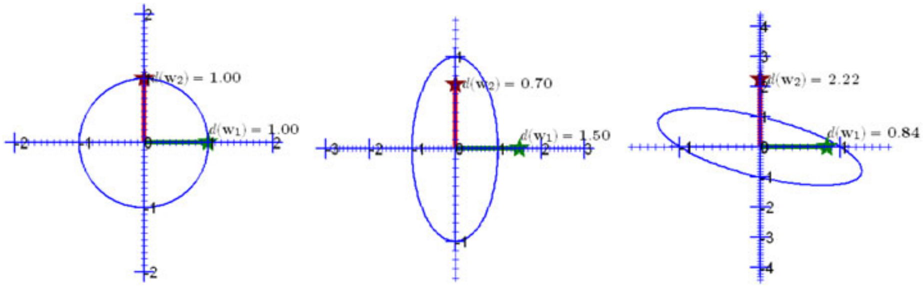
Definition 2 (Scale) Given any translation invariant metric $d:\mathbb{X}\times\mathbb{X}\to\mathbb{R}_+$ and a unit vector $\mathbf{w}\in\mathbb{X}$ where $\|\mathbf{w}\|=1$, the squared distance $d^2(\mathbf{w},\mathbf{0})\triangleq d^2(\mathbf{w})$ is defined as the *scale* of d in direction $\mathbf{w}$.

Definition 3 (Scale extractor & scale vector) Define the operator $\rho_W:\mathcal{F}_{\mathbb{X}}\to\mathbb{R}^n$ which transforms a metric d to a vector consisting of the scales of d on a group of n vectors $W_{m\times n}=[\mathbf{w}_1\ \mathbf{w}_2\ \dots\ \mathbf{w}_n]$ as *scale extractor*:

$$\rho_W(d)=\begin{bmatrix}\rho_{\mathbf{w}_1}(d) & \rho_{\mathbf{w}_2}(d) & \dots & \rho_{\mathbf{w}_n}(d)\end{bmatrix}^\top=\begin{bmatrix}d^2(\mathbf{w}_1) & d^2(\mathbf{w}_2) & \dots & d^2(\mathbf{w}_n)\end{bmatrix}^\top$$

The vector $\rho_W(d)$ is called the *scale vector* of d on W. For Mahalanobis metric d_A, it simply equals to $\rho_W(d_A)=[\mathbf{w}_1^\top A\mathbf{w}_1\ \mathbf{w}_2^\top A\mathbf{w}_2\ \dots\ \mathbf{w}_n^\top A\mathbf{w}_n]^\top$, and we can denote it as $\rho_W(A)$ or ρ_W^A for brevity.

Imagine that a set of unit vectors $\{\mathbf{w}\}_{i=1}^n$ are measured by the "rulers" defined by d_A, and then all these squared distances compose the scale vector $\rho_W(d_A)$. With any fixed W, $\rho_W(d_A)$ is determined by the metric d_A and reflects how the information in these directions is amplified or squashed.



(a) $\rho_W(d_A) = (1.00^2\ 1.00^2)^\top$ (b) $\rho_W(d_A) = (1.50^2\ 0.70^2)^\top$ (c) $\rho_W(d_A) = (2.22^2\ 0.84^2)^\top$

Fig. 2 Examples showing the scale of three different metrics with the same basis $W = [\mathbf{w}_1\ \mathbf{w}_2]$ where $\mathbf{w}_1 = [1\ 0]^\top$ and $\mathbf{w}_2 = [0\ 1]^\top$

This attitude is illustrated in Fig. 2 where the scales in two directions $\mathbf{w}_1 = [1\ 0]^\top$ and $\mathbf{w}_2 = [0\ 1]^\top$ are extracted. We always present the two unit vectors $\mathbf{w}_1, \mathbf{w}_2$ (starting from the origin and ending with a pentagram ★) in the original space, and use an ellipse to show the metric.[5] The ellipse contains all the points with unit distance to the origin measured by d_A, i.e., $\{\mathbf{x} \mid d_A(\mathbf{x}, \mathbf{0}) = 1\}$. Two rulers corresponding to $\mathbf{w}_1$ and $\mathbf{w}_2$ are presented to show the scale properties of d_A in these two directions. If the distance is amplified in one direction, the scale of the ruler becomes denser, such as $\mathbf{w}_1$ in Fig. 2(b) and $\mathbf{w}_2$ in Fig. 2(c). In contrast, the scale of the ruler in the squashed direction becomes sparser. The distances of $\mathbf{w}_1$ and $\mathbf{w}_2$ can then be read directly on the rulers and they compose the scale vector $\rho_W(d_A)$.

In this example, the standard basis of $\mathbb{X}$ is chosen for W. In Fig. 2(a), the distances are measured by Euclidean metric and thus the points with unit distance to the origin simply compose a circle. The scale of a unit vector in any direction is 1. In Fig. 2(b), the Mahalanobis matrix is diagonal. Therefore, its eigenvectors are just the standard basis $\mathbf{w}_1, \mathbf{w}_2$ and the ellipse with unit distance is symmetric with respect to the coordinate axes. Furthermore, $\mathbf{w}_1, \mathbf{w}_2$ correspond to the mostly amplified direction and the mostly squashed direction respectively. In Fig. 2(c), there is no special property of the metric and we just show the scales in the two directions.

Since the scale vector characterizes the most important properties of a metric, it can help to make a study on the metric. This idea is straightforward. Supposing that we are going to measure the shape of a rapidly spinning object, it's neither possible nor necessary to measure directly on its body, but we can take photos of it and measure on the photos instead. In our problem, the metric is like the spinning object which we focus on but is difficult to measure directly. Then the scale extractor plays the role of a camera which takes photos of it. Each scale is one photo characterizing its property from a specific view and the scale vector is the album consisting of all these photos. Furthermore, if we want to compare two spinning objects that are difficult to measure directly, we can resort to their photos instead. Obviously, if the photos of two objects are similar from various views, we can consider them to be similar. Thus, the similarity between d_A and d_B can be measured by $\rho_W(d_A)$ and $\rho_W(d_B)$. In next subsections, utilizing the scale extractor, we will show that a higher geometry preserving probability $\mathrm{PG}_f(d_A, d_B)$ is encouraged by minimizing the von Neumann divergence $D_{\mathrm{vN}}(A, B)$.

[5] We use a different method from Fig. 1 to visualize a Mahalanobis metric.

5.2 Preserving the geometry by minimizing a function of scale vectors

We have shown in Sect. 4.2 that the geometry preserving property is mathematically measured by geometry preserving probability (PG). In this subsection, we show that a higher PG usually accompanies with a smaller $\mathcal{E}(A,B)$, which is a integration defined with the scale vectors. This result transforms the optimization on the complicatedly defined geometry preserving probability into a tractable optimization on a formula of the scale vectors.

Theorem 3 (Geometry Preserving Theorem) *Suppose that there are two pairs of random points* $\mathbf{x}_1,\mathbf{y}_1\in\mathbb{R}^m$ *and* $\mathbf{x}_2,\mathbf{y}_2\in\mathbb{R}^m$ *following some distribution* $f(\mathbf{x}_1,\mathbf{y}_1,\mathbf{x}_2,\mathbf{y}_2)$, *the* geometry preserving probability $\mathrm{PG}_f(d_A,d_B)$ *can be formulated as an integration of a function* $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ *as*

$$\mathrm{PG}_f(d_A,d_B)=\iint_{\mathbb{S}^{m-1}\times\mathbb{S}^{m-1}} R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)\mathrm{d}\Omega(\mathbf{q}_1)\mathrm{d}\Omega(\mathbf{q}_2) \tag{6}$$

where $\mathrm{d}\Omega(\mathbf{q}_i)$ *is the* solid angle element[6] *corresponding to the direction of* $\mathbf{q}_i$ *which contains all the angular factors, and* $\mathbb{S}^{m-1}=\{\mathbf{x}\in\mathbb{R}^m\mid\|\mathbf{x}\|=1\}$ *is the* $(m-1)$*-dimensional unit sphere in* $\mathbb{R}^m$. *The integration is calculated on* $\mathbb{S}^{m-1}$ *for both* $\mathbf{q}_1$ *and* $\mathbf{q}_2$.

Particularly, define

$$\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)=\arctan\sqrt{\frac{\rho^A_{\mathbf{q}_2}}{\rho^A_{\mathbf{q}_1}}}-\arctan\sqrt{\frac{\rho^B_{\mathbf{q}_2}}{\rho^B_{\mathbf{q}_1}}} \tag{7}$$

and assume d_B *(or* d_A*) is given. Then for* $\forall\mathbf{q}_1,\mathbf{q}_2$, *both the* $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ *and* $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$ *are functions of* A *(or* B*) and* $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ *always* decreases *with* $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$:

$$\begin{aligned}\left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A_1,B)\right|<(>)\left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A_2,B)\right|&\Rightarrow R_{\mathbf{q}_1,\mathbf{q}_2}(A_1,B)\ge(\le)R_{\mathbf{q}_1,\mathbf{q}_2}(A_2,B),\\ \left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B_1)\right|<(>)\left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B_2)\right|&\Rightarrow R_{\mathbf{q}_1,\mathbf{q}_2}(A,B_1)\ge(\le)R_{\mathbf{q}_1,\mathbf{q}_2}(A,B_2).\end{aligned}$$

The proof of Theorem 3 is presented in Appendix A.1 for clarity.

By Theorem 3, $\mathrm{PG}_f(d_A,d_B)$ equals to an integration of $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)\mathrm{d}\Omega(\mathbf{q}_1)\mathrm{d}\Omega(\mathbf{q}_2)$, which reflects the geometry preserving property for all pairs $(\mathbf{x}_1,\mathbf{y}_1)$ and $(\mathbf{x}_2,\mathbf{y}_2)$ satisfying $\mathbf{x}_1-\mathbf{y}_1=r_1\mathbf{q}_1$ and $\mathbf{x}_2-\mathbf{y}_2=r_2\mathbf{q}_2$. In order to obtain a higher $\mathrm{PG}_f(d_A,d_B)$, we have to solve A and B that maximize the integration of $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ shown in (6) and satisfy the constraints from side-information. It is difficult to give a precise analysis in the general case because $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ is influenced by the distribution f, which is indeterminate. However, no matter what the distribution f is, $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ always monotonically decreases with $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$. Thus, considering that we just want to maximize (6) rather than to obtain its exact value, in general, we can approximately achieve this goal by replacing $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ in (6) with $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$ and then minimizing the integration (8) instead.

$$\mathcal{E}(A,B)=\iint_{\mathbb{S}^{m-1}\times\mathbb{S}^{m-1}}\left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)\right|\mathrm{d}\Omega(\mathbf{q}_1)\mathrm{d}\Omega(\mathbf{q}_2). \tag{8}$$

[6] Considering an $(m-1)$-dimensional sphere in $\mathbb{R}^m$ with radius r, the *solid angle element* $\mathrm{d}\Omega(\mathbf{q}_i)$ is the corresponding surface element divided by r^{m-1}, which numerically equals to the surface element on a unit sphere. In the case of $m=2$, it degenerates to a common angle element. Please refer to Appendix A.1 for the definition of $\mathrm{d}\Omega(\mathbf{q}_i)$ in detail.

Due to the nonnegativity of $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$, if $\mathcal{E}(A,B)=0$, it is obvious that $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|=0$, $\forall \mathbf{q}_1,\mathbf{q}_2$ and $\mathrm{PG}_f(d_A,d_B)$ reaches the maximum 1. Along with the increase of $\mathcal{E}(A,B)$, the $\mathrm{PG}_f(d_A,d_B)$ begins to decrease. Therefore, $\mathcal{E}(A,B)$ is a generic approximation of the deterioration of $\mathrm{PG}_f(d_A,d_B)$ without knowledge of f, and minimizing $\mathcal{E}(A,B)$ has the effect to increase $\mathrm{PG}_f(d_A,d_B)$.

5.3 Bounding the Bregman divergence of scales by Bregman matrix divergence

In this section, we show how two Mahalanobis metrics d_A and d_B are enforced to be correlated with each other by minimizing the Bregman matrix divergence $D_\phi(A,B)$. As two functions, the relationship between d_A and d_B is not explicit from $D_\phi(A,B)$. Since the scale vector reveals the important scale properties of one Mahalanobis metric and is much simpler to deal with, we resort to ρ_W^A and ρ_W^B to investigate the relationship between d_A and d_B.

Bregman divergence (Dhillon and Tropp 2008) is a class of widely-used diversity measures for vectors in machine learning, including the *squared Euclidean distance*, *generalized KL-divergence*, *Itakura-Saito distance*, etc. For $\forall \mathbf{x},\mathbf{y}\in\mathbb{R}^m$, it is defined as

$$D_\varphi(\mathbf{x},\mathbf{y})=\varphi(\mathbf{x})-\varphi(\mathbf{y})-\nabla\varphi(\mathbf{y})^\top(\mathbf{x}-\mathbf{y}),$$

where $\varphi(\cdot)$ is a convex *generating function* defined on $\mathbb{R}^m$. If $\varphi(\mathbf{x})=\mathbf{x}^\top\mathbf{x}$, we obtain the squared *Euclidean distance* $D_\varphi(\mathbf{x},\mathbf{y})=\|\mathbf{x}-\mathbf{y}\|^2$; if $\varphi(\mathbf{x})=\sum_i x_i\log x_i - x_i$, we obtain the *KL-divergence* $D_{\mathrm{KL}}(\mathbf{x},\mathbf{y})=\sum_i x_i(\log x_i-\log y_i)-x_i+y_i$.

Compared with the definition of Bregman matrix divergence (2), the Bregman divergence has an identical form except that it takes real vectors as variable instead of Hermitian matrices. By Corollary 2 of Kulis et al. (2009), there is the relationship between them as

$$D_{\varphi\circ\lambda}(X,Y)=\sum_{i,j}\left(\mathbf{v}_i^\top\mathbf{u}_j\right)^2 D_\varphi(\lambda_i,\theta_j), \tag{9}$$

where $X=\sum_i\lambda_i\mathbf{v}_i\mathbf{v}_i^\top$, $Y=\sum_i\theta_i\mathbf{u}_i\mathbf{u}_i^\top$ are spectral decompositions, and thus $\{\mathbf{v}_i\}_{i=1}^m$, $\{\mathbf{u}_i\}_{i=1}^m$ are both orthonormal bases. Since $V^\top U=[\mathbf{v}_i^\top\mathbf{u}_j]_{m\times m}$ is orthogonal, the matrix $[(\mathbf{v}_i^\top\mathbf{u}_j)^2]_{m\times m}$ is *orthostochastic* and thus *doubly stochastic*, whose row and column sums are 1 (Horn and Johnson 1991). Therefore, the matrix divergence is regarded as the sum of scalar divergences between pairs of eigenvalues, weighted by the squared cosine of the angle between the corresponding eigenvectors (Dhillon and Tropp 2008).

Among the Bregman matrix divergences, the $\|A-B\|_\mathrm{F}^2$ and $D_{\mathrm{vN}}(A,B)$ are specifically appropriate for the framework (1) due to the following reasons:

1. Both of them are jointly convex with respect to the two arguments, which guarantees a global optimal solution as long as the loss function determined by the metric learning algorithm is convex. The joint convexity of $\|A-B\|_\mathrm{F}^2$ is straightforward, while the joint convexity of $D_{\mathrm{vN}}(A,B)$ is presented in Theorem 1.
2. Both of them provide a bound for its corresponding Bregman divergence of the scale vectors of the metrics, which links the Bregman matrix divergence of two Mahalanobis matrices with the Bregman divergence of their scales.[7] Specifically, if $\varphi(\mathbf{x})=\|\mathbf{x}\|^2$ or

[7] Unfortunately, this result cannot be straightforwardly extended to other Bregman divergences. By numerical experiments, we found that the inequality does not hold for $\varphi(\mathbf{x})=-\sum_i\log x_i$ which corresponds to the LogDet divergence and the Itakura-Saito divergence (Dhillon and Tropp 2008).

$\varphi(\mathbf{x}) = \sum_i \log x_i - x_i$, then for any orthonormal bases $W = [\mathbf{w}_1 \ \dots \ \mathbf{w}_m]$, we have

$$D_\varphi\big(\rho_W^A, \rho_W^B\big) \le D_\phi(A, B), \tag{10}$$

where $\phi = \varphi \circ \lambda$ and $\lambda(A)$ is the function that lists the eigenvalues of A. This result is formally presented in Theorem 4.

Theorem 4 *Suppose $d_A, d_B \in \mathcal{F}_{\mathbb{R}^m}$ are two Mahalanobis metrics defined on $\mathbb{R}^m$. Then for any orthonormal basis $W = [\mathbf{w}_1 \ \dots \ \mathbf{w}_m]$ in $\mathbb{R}^m$, the* squared Frobenius norm of the difference *and the* von Neumann divergence *of their Mahalanobis matrices A and B provides an upper bound for the* squared Euclidean distance *and the* KL-divergence *of their scale vectors ρ_W^A and ρ_W^B respectively*:

$$\|A - B\|_{\mathrm{F}}^2 = \sup_{W^\top W = \mathbf{I}_m} \big\|\rho_W^A - \rho_W^B\big\|^2, \tag{11}$$

$$D_{\mathrm{vN}}(A, B) \ge \sup_{W^\top W = \mathbf{I}_m} D_{\mathrm{KL}}\big(\rho_W^A, \rho_W^B\big). \tag{12}$$

In spite of their uniform formulation (10), the two cases resort to very different proofs, and we leave them in Appendix A.2.

Recall the example we presented at the end of Sect. 5.1 where the discrepancy between the shapes of two spinning objects (metrics) are measured by comparing their photos (scale vectors), then ρ_W with an orthogonal W determines a minimal set of cameras that can cover the complete views. Each camera $\rho_{\mathbf{w}_i}$ takes a photo for d_A, d_B respectively and their discrepancy from this view is measured by $D_\varphi(\rho_{\mathbf{w}_i}^A, \rho_{\mathbf{w}_i}^B)$. Then $D_\varphi(\rho_W^A, \rho_W^B)$ gives the discrepancies of d_A and d_B by summing up the results from all cameras. Theorem 4 provides an upper bound for such a total measure on various views, which captures the overall discrepancy of the two objects (metrics).

Furthermore, define the continuous version of $D_\varphi(\rho_W^A, \rho_W^B) = \sum_i D_\varphi(\rho_{\mathbf{w}_i}^A, \rho_{\mathbf{w}_i}^B)$ as

$$\mathcal{D}_\varphi(A, B) = \int_{\mathbb{S}^{m-1}} D_\varphi\big(\rho_{\mathbf{q}}^A, \rho_{\mathbf{q}}^B\big) \mathrm{d}\Omega(\mathbf{q}), \tag{13}$$

which summarizes the discrepancy of two metrics measured in all directions. Denote $\mathcal{D}_\varphi$ corresponding to $\|\rho_{\mathbf{q}}^A - \rho_{\mathbf{q}}^B\|^2$ and $D_{\mathrm{KL}}(\rho_{\mathbf{q}}^A, \rho_{\mathbf{q}}^B)$ as $\mathcal{D}_{\mathrm{Eu}}$ and $\mathcal{D}_{\mathrm{KL}}$ respectively. Since $D_\varphi(\rho_W^A, \rho_W^B) \le D_\phi(A, B)$ holds for any orthogonal W, minimizing $D_\phi(A, B)$ has an effect to minimize $\mathcal{D}_\varphi(A, B)$. This subsection relates $D_\phi(A, B)$ with a discrepancy measure defined by scale vectors, which enables us to further establish the relationship between $\mathrm{PG}_f(d_A, d_B)$ and $D_\phi(A, B)$ as shown in next subsection.

5.4 Preserving geometry property by minimizing von Neumann divergence

In this subsection, based on the results shown in the above subsections, we present our conclusion that a high geometry preserving probability $\mathrm{PG}_f(d_A, d_B)$ is usually better encouraged by minimizing $D_{\mathrm{vN}}(A, B)$ than by minimizing $\|A - B\|_{\mathrm{F}}^2$ for multi-task problems.

For convenience of explanation, we first informally define the concept of *consistent*. Assuming that there are two functions $f(\mathbf{x}), g(\mathbf{x})$, if for randomly selected $\mathbf{x}_1, \mathbf{x}_2 \in \mathrm{dom} f \cap \mathrm{dom} g$, the assertion $f(\mathbf{x}_1) > f(\mathbf{x}_2) \Leftrightarrow g(\mathbf{x}_1) > g(\mathbf{x}_2)$ holds with a high probability, we call that the two functions are *consistent*. It is obvious that if two functions are consistent, each of them is likely to decrease or increase with the other one and thus minimizing either is to minimize the other. Then we establish the relationship between $D_\phi(A,B)$ and $\mathrm{PG}_f(d_A,d_B)$ in the following steps:

1. Section 5.2 explains that a higher $\mathrm{PG}_f(d_A,d_B)$ usually accompanies with a smaller $\mathcal{E}(A,B)$, and thus a better geometry preserving property can be encouraged by minimizing $\mathcal{E}(A,B)$. We denote this fact as

$$\mathrm{PG}_f(d_A,d_B)\uparrow \quad \Leftrightarrow \quad \mathcal{E}(A,B)\downarrow. \tag{14}$$

2. Section 5.3 proves that the regularization to be minimized $D_\phi(A,B)$ provides an upper bound for $D_\varphi(\rho_W^A,\rho_W^B)$, which furthermore implies that $\mathcal{D}_\varphi(A,B)$ is minimized in our framework. We denote this fact as

$$D_\phi(A,B)\downarrow \Rightarrow \mathcal{D}_\varphi(A,B)\downarrow. \tag{15}$$

3. Our object is to bridge the left side of (14) and (15), while the right side of them, i.e. $\mathcal{E}(A,B)$ and $\mathcal{D}_\varphi(A,B)$, are both integrations of scale vectors. Thus it provides a way to bridge $D_\phi(A,B)$ and $\mathrm{PG}_f(d_A,d_B)$. Specifically, if there is a type of $\mathcal{D}_\varphi(A,B)$ consistent with $\mathcal{E}(A,B)$, i.e., $\mathcal{D}_\varphi(A,B)\downarrow \Leftrightarrow \mathcal{E}(A,B)\downarrow$, we can obtain that

$$D_\phi(A,B)\downarrow \Rightarrow \mathcal{D}_\varphi(A,B)\downarrow \quad \Leftrightarrow \quad \mathcal{E}(A,B)\downarrow \quad \Leftrightarrow \quad \mathrm{PG}_f(d_A,d_B)\uparrow,$$

which means a good geometry preserving property can be obtained by minimizing the corresponding Bregman matrix divergence.

In this subsection, we will show that $\mathcal{D}_{\mathrm{KL}}(A,B)$ is *more consistent* with $\mathcal{E}(A,B)$ compared with $\mathcal{D}_{\mathrm{Eu}}(A,B)$, which proves $D_{\mathrm{vN}}(A,B)$ a better candidate of the regularization to preserve geometry between metrics.

In multi-task problems, the difference between any two metrics is usually relatively small, and we investigate the consistency of $\mathcal{E}(A,B)$ and $\mathcal{D}_\varphi(A,B)$ based on this assumption. When $d_A=d_B$, all scales are equal as $\rho_{\mathbf{q}}^A\equiv\rho_{\mathbf{q}}^B, \forall\mathbf{q}$ and thus $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|\equiv 0 \Rightarrow \mathcal{E}(A,B)=\mathcal{D}_\varphi(A,B)=0$. As d_A becomes different from d_B so that there is a difference of scales in some direction $\rho_{\mathbf{q}}^A-\rho_{\mathbf{q}}^B$, both $\mathcal{E}(A,B)$ and $\mathcal{D}_\varphi(A,B)$ will thus increase.

If $\mathcal{E}(A,B)$ and $\mathcal{D}_\varphi(A,B)$ are consistent, a scale difference that brings about a greater increment of $\mathcal{E}(A,B)$ is also expected to produce a greater increment of $\mathcal{D}_\varphi(A,B)$, and vice versa. The increments are determined by both the value of scale difference $\rho_{\mathbf{q}}^A-\rho_{\mathbf{q}}^B$ and the direction $\mathbf{q}$. In the same direction $\mathbf{q}$, it's easy to see that both the two functions increase with the absolute difference of scales $|\rho_{\mathbf{q}}^A-\rho_{\mathbf{q}}^B|$ and keep consistent. In the following, we investigate the increments of the two functions for scale differences in different directions.

First we study how $\mathcal{E}(A,B)$ is influenced by the difference in each direction and the result is presented in Proposition 1.

Proposition 1 *Assume that there are three Mahalanobis metrics $d_{A_1}, d_{A_2}, d_B \in \mathcal{F}_{\mathbb{R}^m}$, and the unit vectors $\mathbf{w}_1, \mathbf{w}_2$ define two directions. Extracting the scales of the metrics by ρ_W, we obtain that d_{A_1} and d_{A_2} differ from d_B on $\mathbf{w}_1$ and $\mathbf{w}_2$ respectively as $\rho_{\mathbf{w}_1}^{A_1}-\rho_{\mathbf{w}_1}^B=\rho_{\mathbf{w}_2}^{A_2}-\rho_{\mathbf{w}_2}^B=\Delta\rho$. If the difference $\Delta\rho$ is relatively small compared to $\rho_{\mathbf{w}_i}$, we have*

$$\frac{\mathcal{E}(A_1,B)}{\mathcal{E}(A_2,B)}\approx\left(\frac{\rho_{\mathbf{w}_2}^B}{\rho_{\mathbf{w}_1}^B}\right)^\alpha, \tag{16}$$

where $0.5<\alpha<1.5$.

The proof of Proposition 1 is presented in Appendix A.3 for clarity.

From Proposition 1, the increase of $\mathcal{E}(A, B)$ brought about by $\rho^A_{\mathbf{w}_i} - \rho^B_{\mathbf{w}_i}$ is approximately inversely proportional to $\rho^B_{\mathbf{w}_i}$, and thus the deterioration of the geometry preserving property is determined by the *relative variation* of the scale. This coincides with our intuition in that the geometrical property is more sensitive to the variation of the smaller scale, which is also a result of the fact that the scale essentially defines a squared *magnification factor* for the distance. Assuming that $\rho^B_{\mathbf{w}_1} = 10$, $\rho^B_{\mathbf{w}_2} = 0.1$, then if $\rho^B_{\mathbf{w}_1}$ is increased by 1, the distance of any pair of points in direction $\mathbf{w}_1$ is magnified to $\sqrt{11}/\sqrt{10} = 1.05$ times, while if $\rho^B_{\mathbf{w}_2}$ is increased by 1, the distance of any pair of points in direction $\mathbf{w}_2$ is magnified to $\sqrt{1.1}/\sqrt{0.1} = 3.32$ times.

To be a discrepancy measure consistent with $\mathcal{E}(A, B)$, a proper $\mathcal{D}_\varphi(A, B)$ should also be more sensitive to the difference of the smaller scale. Then we analyze how the two types of $\mathcal{D}_\varphi(A, B)$ increase with the difference of scales and present the results in Proposition 2.

Proposition 2 *Assume that all the variables are identically defined as in Proposition* 1, *we have*

$$\frac{\mathcal{D}_{\mathrm{Eu}}(A_1, B)}{\mathcal{D}_{\mathrm{Eu}}(A_2, B)} = 1, \qquad \frac{\mathcal{D}_{\mathrm{KL}}(A_1, B)}{\mathcal{D}_{\mathrm{KL}}(A_2, B)} \approx \frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}. \tag{17}$$

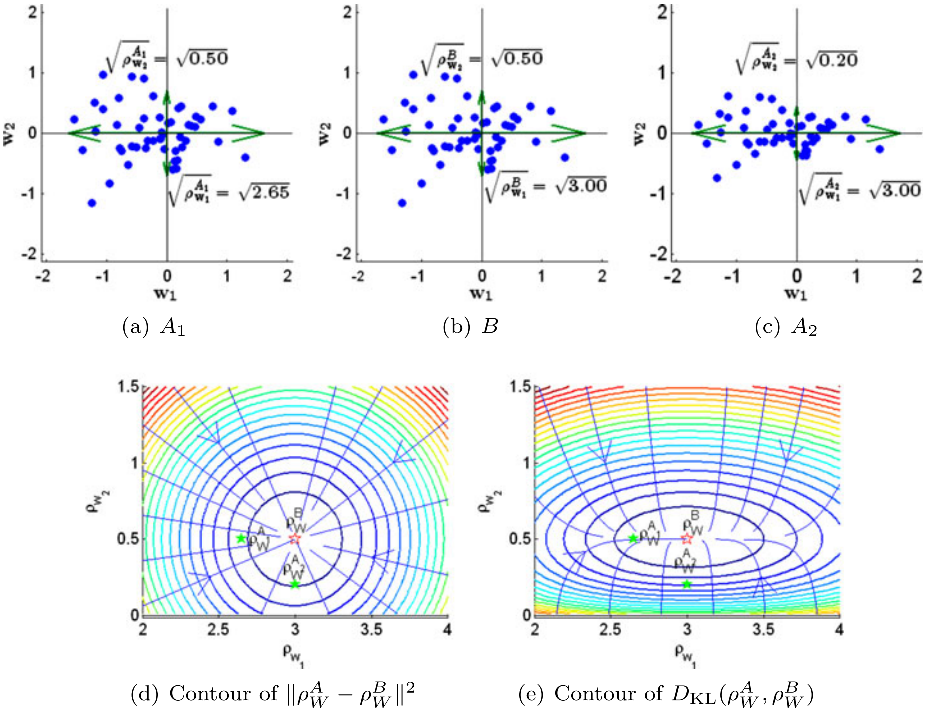
We also present the proof of Proposition 2 in Appendix A.3.

If we compare (17) with (16), both $\mathcal{D}_{\mathrm{KL}}(A, B)$ and $\mathcal{E}(A, B)$ are more sensitive to the difference of the smaller scale, where the increments of them brought about by $\rho^A_{\mathbf{w}_i} - \rho^B_{\mathbf{w}_i}$ are both approximately inversely proportional to $\rho^B_{\mathbf{w}_i}$. In contrast, the increment of $\mathcal{D}_{\mathrm{Eu}}(A, B)$ is independent with the direction or scale but determined by only the value of scale difference. Thus the increment of $\mathcal{D}_{\mathrm{KL}}(A, B)$ is more consistent with $\mathcal{E}(A, B)$ compared with $\mathcal{D}_{\mathrm{Eu}}(A, B)$, which implies that if $\mathcal{D}_{\mathrm{KL}}(A_1, B) > \mathcal{D}_{\mathrm{KL}}(A_2, B)$, it is more likely to obtain that $\mathcal{E}(A_1, B) > \mathcal{E}(A_2, B)$ while $\mathcal{D}_{\mathrm{Eu}}$ does not have such a property.

It's notable that the result of Proposition 1 is obtained by considering only the variation of one scale and does not precisely hold when several scales change simultaneously because the partial derivative in (16) becomes more complex. However, in most cases, the conclusion that $\mathcal{E}(A, B)$ is more sensitive to the variation of the smaller scale still holds and thus $\mathcal{D}_{\mathrm{KL}}(A, B)$ reflects the deterioration of PG more accurately. Then from the discussion of the beginning of this subsection, $D_{\mathrm{vN}}(A, B)$ is a better choice of the regularization to preserve the geometry.

This phenomenon is illustrated by the example in Fig. 3. The scale vectors of the three metrics measured in the directions of standard basis are $\rho^B_W = [3.00\ 0.50]^\top$, $\rho^{A_1}_W = [2.65\ 0.50]^\top$, and $\rho^{A_2}_W = [3.00\ 0.20]^\top$, where d_{A_1} and d_{A_2} differ from d_B on $\mathbf{w}_1$ and $\mathbf{w}_2$ respectively. Since $\rho^B_{\mathbf{w}_1} > \rho^B_{\mathbf{w}_2}$, the metric is more sensitive to the difference in direction $\mathbf{w}_2$ and PG decreases more rapidly in this direction. Therefore, even though the scale difference between d_{A_1} and d_B is greater as $|\rho^{A_1}_{\mathbf{w}_1} - \rho^B_{\mathbf{w}_1}| = 0.35 > |\rho^{A_2}_{\mathbf{w}_2} - \rho^B_{\mathbf{w}_2}| = 0.30$, the geometry of samples measured by d_{A_1} look more similar to d_B compared with d_{A_2}, This phenomenon can be explained as that d_{A_1} preserves the geometry (relative distances) from d_B better than d_{A_2} does, which can be verified by comparing the geometry preserving probabilities. Estimating the PG using randomly generated samples,[8] we obtain $\mathrm{PG}_f(A_1, B) \approx 0.990 > \mathrm{PG}_f(A_2, B) \approx 0.936$, which confirms our conjecture above.

[8] Please refer to Appendix A.4 for the detail of the procedure.



(a) A_1 (b) B (c) A_2

(d) Contour of $\|\rho^A_W - \rho^B_W\|^2$ (e) Contour of $D_{\mathrm{KL}}(\rho^A_W, \rho^B_W)$

Fig. 3 The geometry property of d_A and d_B is better preserved by minimizing $D_{\mathrm{vN}}(A, B)$ rather than $\|A - B\|^2_{\mathrm{F}}$. On one hand, with the randomly generated samples, the geometry probabilities are estimated to be $\mathrm{PG}_f(A_1, B) \approx 0.990 > \mathrm{PG}_f(A_2, B) \approx 0.936$ and thus A_1 preserved the geometry property from B better than A_2 does. On the other hand, it is straightforward to calculate that $\|A_1 - B\|^2_{\mathrm{F}} = 0.1225 > \|A_2 - B\|^2_{\mathrm{F}} = 0.0900$, and $D_{\mathrm{vN}}(A_1, B) = 0.0213 < D_{\mathrm{vN}}(A_2, B) = 0.1167$. Therefore, von Neumann divergence correctly assigns a lower discrepancy measure to A_1 which preserves the geometry properties from B better, and thus minimizing $D_{\mathrm{vN}}(A, B)$ prefers A_1 to A_2 as the metric similar to B. Furthermore, from the contours of $D_\varphi(\rho^A_W, \rho^B_W)$ with respect to ρ^A_W, we see that the $D_{\mathrm{KL}}(\rho^A_W, \rho^B_W)$ corresponding to $D_{\mathrm{vN}}(A, B)$ increases more rapidly in the direction with respect to the smaller scale

On the other hand, calculating the Bregman matrix divergences, we obtain that $\|A_1 - B\|^2_{\mathrm{F}} = 0.1225 > \|A_2 - B\|_{\mathrm{F}} = 0.0900$ and $D_{\mathrm{vN}}(A_1, B) = 0.0213 < D_{\mathrm{vN}}(A_2, B) = 0.1167$. Obviously, von Neumann divergence provides a discrepancy measure that is more consistent with the deterioration of geometry preserving property. Suppose that we want to encourage d_A to be similar to d_B by minimizing $D_\phi(A, B)$, then d_{A_1} is preferred to d_{A_2} if $D_\phi(A, B) = D_{\mathrm{vN}}(A, B)$, while d_{A_2} is preferred to d_{A_1} if $D_\phi(A, B) = \|A - B\|^2_{\mathrm{F}}$. Therefore, von Neumann divergence can select the correct one with the better geometry preserving property.

Besides, when the squared Frobenius norm is used, the obtained solution may have negative eigenvalues and we have to project the solution to the positive semi-definite cone. Instead, von Neumann divergence can automatically keep the solution to be positive semi-definite.

In summary, minimizing the von Neumann divergence usually encourages a higher geometry preserving probability in multi-task problems, and thus it is a more appropriate regularization to couple different metrics. Using $D_{\mathrm{vN}}(A, B)$ as the regularization in (1), it is expected to obtain a better geometry preserving property.

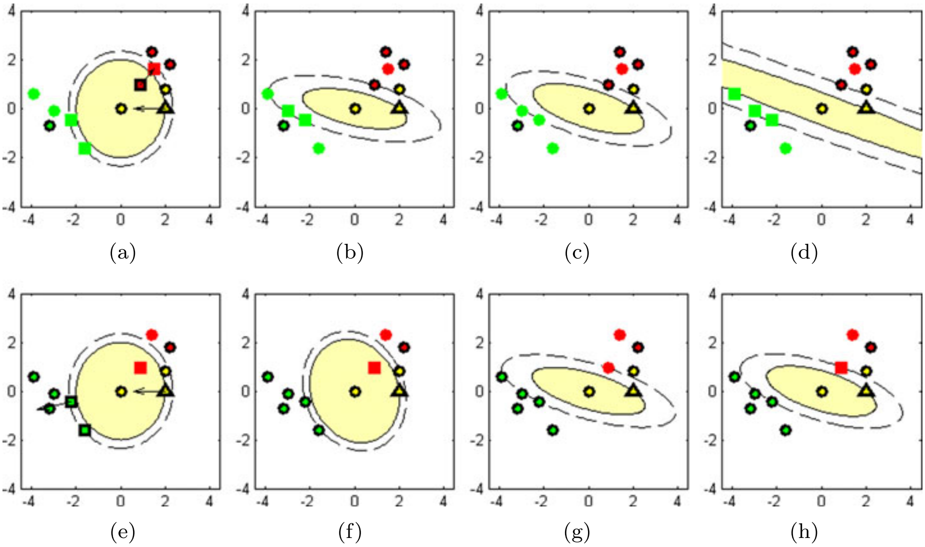


Fig. 4 An illustration of multi-task metric learning. (**a**), (**e**) The original data of task 1/2. (**b**), (**f**) The data of task 1/2 after single task metric learning. (**c**), (**g**) The data of task 1/2 after joint metric learning using von Neumann divergence as regularization. (**d**), (**h**) The data of task 1/2 after joint metric learning using squared Frobenius norm of difference as regularization. Joint learning of multiple tasks (given by our proposed geometry preserving framework) can lead to ideal metrics for both task-1 in (**c**) & task-2 in (**g**) (Color figure online)

6 Experiments

6.1 A toy example

In this section, we use a toy example in Fig. 4 to show the advantage of the von Neumann divergence in preserving the geometry relationship between samples. There are two related classification tasks, each of which has 3 classes and the samples are shown in Figs. 4(a) and 4(e) respectively. The color of each point indicates its label and the shape represents its role in metric learning which we will explain later. A point with a black border represents a training sample while the one without a border represents a test sample. Unfortunately, in training set, there is only one green point for task-1 and one red point for task-2, which cannot represent the distribution of the corresponding class accurately. Observing that the samples of two tasks have very similar distributions, we are motivated to utilize the information from the training samples of the other task to improve the performance for both of the tasks.

Here we use the idea of LMNN to learn a better metric. Focusing on the yellow point in the center of the figure, LMNN aims to learn a metric so that the nearest neighbor of this point belongs to the same class. To obtain such a metric, the nearest yellow point is chosen as the *target* point (represented with △) and a circle through this point is drawn. Then any point belonging to a different class is expected to be further than any target with a large margin and thus stand outside the dashed perimeter. Any point with a different label lying inside the dashed perimeter is called *imposter* (represented with ■) and the objective of LMNN is to pull the target closer while pushing all imposters outside the perimeter. This encourages the similar samples to be closer to each other. In Fig. 4, we show the learned metric by drawing an ellipse formed by the locus of points equidistant from the central point (the same way as Fig. 2). Then the distance from any point on the dashed ellipse to the central point equals to the distance from the target to the center, plus a margin. Thus any red or green point lying inside the perimeter is an imposter and should be pushed out.

Figure 4(b) shows the learned metric of task-1, where all the red imposters (both training and test points) are pushed outside while Fig. 4(f) shows task-2 with all green imposters outside. Unfortunately, for task-1, since the green points in training set are too few to represent the distribution, the learned metric cannot accurately separate the green class from the yellow one and some test samples invade the perimeter. The same situation also happens for the red class in task-2.

Since the distribution of two tasks are very similar, we hope to let task-1 borrow information about the green class from task-2, and task-2 borrow information about the red class from task-1. Denote the Mahalanobis matrices with respect to Figs. 4(f) and 4(b) as A_1 and A_2. We have $A_1 \in \mathcal{C}_1$ and $A_2 \in \mathcal{C}_2$ since they satisfy the constraints from side-information of task-1 and task-2 respectively. Recall that we propagate information from A to B by minimizing $D(A, B)$ and solve a better metric for task-1 by

$$\min_{A \in \mathcal{C}_1} D(A, A_2). \tag{18}$$

Since $A_2 \in \mathcal{C}_2$, it is equivalent to solving a metric which satisfies all constraints from $\mathcal{C}_1$ and as many constraints from $\mathcal{C}_2$ as possible. In this example, it should push the red imposter in Fig. 4(f) out of the perimeter while trying the best to keep the green points outside. The problem is defined in the same way for task-2.

The solutions of (18) using $D(A, B) = D_{\mathrm{vN}}(A, B)$ and $D(A, B) = \|A - B\|_{\mathrm{F}}^2$ are shown in Figs. 4(c) and 4(d) respectively. From the figures, we see that if von Neumann divergence is used, constraints of both tasks are satisfied by the learned metric, i.e., $A \in \mathcal{C}_1 \cap \mathcal{C}_2$. There is no imposter in either training set or test set, and both the red and green classes are separated well. This shows that the geometry property of samples is preserved from task-2 to task-1 and the side-information of task-2 is well propagated to task-1. In contrast, if the Frobenius norm of difference is used, to push the red imposter outside the perimeter, some green points invade into this perimeter again, which produces more imposters. This is because the geometry property of A_2 that discriminates the green class from the yellow one is not preserved. For the case to improve task-2 with task-1, the results are shown in Figs. 4(g) and 4(h) and von Neumann divergence also performs better than Frobenius norm.

6.2 Experiments on real data

To validate our proposed approach, we apply our multi-task framework to the famous LMNN (Weinberger and Saul 2009) metric learning method and conduct experiments on several real data sets. We have introduced its basic idea in Sect. 6.1 and the loss function is simply the sum of all squared distances between samples and their target neighbors, i.e., $\sum_{i,j \rightsquigarrow i} d_{A_t}^2(\mathbf{x}_{ti}, \mathbf{x}_{tj})$, where $j \rightsquigarrow i$ means $\mathbf{x}_{tj}$ is a target neighbor of $\mathbf{x}_{ti}$. The constraints require all imposters stand further than the target neighbors with a margin, i.e., $d_{A_t}^2(\mathbf{x}_{ti}, \mathbf{x}_{tl}) - d_{A_t}^2(\mathbf{x}_{ti}, \mathbf{x}_{tj}) \geq 1$ for $\forall j \rightsquigarrow i$ and $y_{tl} \neq y_{ti}$ where y_{ti} is the label of the i-th sample of task-t. Since there may be no metric satisfying all constraints, a relaxed version of the constraints are used by introducing slack variables. The obtained loss function for task-t is then

$$(1-\mu) \sum_{i,j \rightsquigarrow i} d_{A_t}^2(\mathbf{x}_{ti}, \mathbf{x}_{tj}) + \mu \sum_{i,j \rightsquigarrow i} \sum_{l} (1 - y_{til}) \big[1 + d_{A_t}^2(\mathbf{x}_{ti}, \mathbf{x}_{tj}) - d_{A_t}^2(\mathbf{x}_{ti}, \mathbf{x}_{tl})\big]_+,$$

where $y_{til} = 1$ if and only if $y_{ti} = y_{tl}$, and $y_{til} = 0$ otherwise, $[z]_+ = \max(z, 0)$. We use the fast solver (Weinberger and Saul 2008) to solve the LMNN and our code is based on the original LMNN code.[9]

Every data set contains several related classification tasks, each of which is to predict the labels of the test samples using their features. For each task, a Mahalanobis metric is learned from the training samples, and then the label of each test sample is predicted by the nearest neighbor classifier, where the distance is calculated using the learned metric.

The multi-task learning setting is categorized into the *label compatible* and *label incompatible* scenarios, according to their label sets (Parameswaran and Weinberger 2010). For label incompatible scenario where the label sets of these tasks are different, we compared our method (*MT von Neumann*) with *Euclidean*, *Single Task*, *mtLMNN*, and *MT Frobenius*. The training samples of each task are used as the prototypes of the nearest neighbor classifier. For label compatible scenario where all tasks share the same label set, we also combined the samples of all tasks and learn a unique metric (*Unique Task*). Besides, for *Unique Task*, *mtLMNN*, *MT Frobenius*, and *MT von Neumann*, we also implement a "pooling" version of testing (with a suffix "-pool" after the name) on each of them, where the training samples of all tasks are used as the prototypes. The details of all the compared methods are shown below with a summary in Table 1.

1. *Euclidean*—The nearest neighbor of each test sample is searched in the training samples of this task where the distance is determined by the Euclidean metric directly.
2. *Single Task*—For each task, a metric is learned individually for each task. Then the classifier finds the nearest neighbor in the training samples set of this task using the learned metric.
3. *Unique Task*—The training samples of all tasks are mixed into one sample set and a unique metric is learned from it. Then the nearest neighbor is found in the training samples of this task using the learned metric.
4. *Unique Task-pool*—The same metric as *Unique Task* is used while the nearest neighbor is searched in the training samples of all the tasks using the learned metric.
5. *mtLMNN*—The method proposed by Parameswaran and Weinberger (2010) which has been introduced in Sect. 3.1. It is the same as *MT Frobenius* approach with an additional constraint[10] $A_t \succeq B \succeq 0$. The nearest neighbor is searched in the training sample of this task using the learned metric.
6. *mtLMNN-pool*—The same metric as *mtLMNN* is used while the nearest neighbor is searched in the training samples of all the tasks using the learned metric.
7. *MT Frobenius*—The framework (1) with $D(A, B) = \|A - B\|_{\mathrm{F}}^2$. As we indicated in Sect. 4.1, the constraint $A_t \succeq B$ in mtLMNN is too strong. By replacing it with $A_t \succeq 0$, the relation of A_t and B is more flexible and expected to perform better. The nearest neighbor is searched in the training samples of this task using the learned metric.
8. *MT Frobenius-pool*—The same metric as *MT Frobenius* is used while the nearest neighbor is searched in the training samples of all the tasks using the learned metric.
9. *MT von Neumann*—Our proposed geometry preserving multi-task metric learning. It is the framework (1) with $D(A, B) = D_{\mathrm{vN}}(A, B)$. The nearest neighbor is searched in the training samples of this task using the learned metric.
10. *MT von Neumann-pool*—The same metric as *MT von Neumann* is used while the nearest neighbor is searched in the training samples of all the tasks using the learned metric.

[9] The code of LMNN is downloadable from http://www.cse.wustl.edu/~kilian/code/code.html.

[10] See Sect. 4.1 for detail.

Table 1 Summary of our compared methods

Method	$D(A_t, B)$	Training	Prototype	Constraints
Euclidean	–	–	$\mathcal{X}_t$	–
Single Task	–	$\mathcal{X}_t$	$\mathcal{X}_t$	$A_t \succeq 0$
Uniform Task	$A_t = B$	$\bigcup \mathcal{X}_t$	$\mathcal{X}_t$	$A_t \succeq 0$
Uniform Task-pool	$A_t = B$	$\bigcup \mathcal{X}_t$	$\bigcup \mathcal{X}_t$	$A_t \succeq 0$
MT Frobenius	$\|A_t - B\|_{\mathrm{F}}^2$	$\mathcal{X}_t$	$\mathcal{X}_t$	$A_t \succeq 0$
MT Frobenius-pool	$\|A_t - B\|_{\mathrm{F}}^2$	$\mathcal{X}_t$	$\bigcup \mathcal{X}_t$	$A_t \succeq 0$
mtLMNN	$\|A_t - B\|_{\mathrm{F}}^2$	$\mathcal{X}_t$	$\mathcal{X}_t$	$A_t \succeq B \succeq 0$
mtLMNN-pool	$\|A_t - B\|_{\mathrm{F}}^2$	$\mathcal{X}_t$	$\bigcup \mathcal{X}_t$	$A_t \succeq B \succeq 0$
MT von Neumann	$D_{\mathrm{vN}}(A_t, B)$	$\mathcal{X}_t$	$\mathcal{X}_t$	$A_t \succeq 0$
MT von Neumann-pool	$D_{\mathrm{vN}}(A_t, B)$	$\mathcal{X}_t$	$\bigcup \mathcal{X}_t$	$A_t \succeq 0$

In the model of LMNN, there are two hyper-parameters: (1) a coefficient to balance the loss function and the soft constraints; (2) the number of targets. To determine them, we perform the 5-folder cross-validation on the single-task LMNN using the training samples, and the optimal parameters are selected according to the average error of all tasks, which are then used for all methods. We do not adjust the hyper-parameters in the model of LMNN for each method but use the same values selected for the single-task approach. There are also two hyper-parameters γ and γ_0 in each model of *mtLMNN*, *MT Frobenius* and *MT von Neumann*. We will show how to select them for each dataset in the following.

6.2.1 Multi-speaker vowel classification

The vowel classification data set consists of 11 vowels uttered by 15 speakers of British English, each vowel is said six times. The speakers are divided into two subgroups according to their gender since men pronounce in a different style from women, and each subgroup is regarded as a task. Then we can utilize the multi-task learning to learn a metric for each task.

Considering that multi-task learning aims to deal with the situation where training samples are insufficient, we randomly select only a portion of samples from the vowels of speakers 1–8 and use them to learn a metric, which is tested on the vowels of speakers 9–15. Since the two tasks share the common label set, all the 10 methods are evaluated on this data set. The optimal hyper-parameters for each method are selected by a 5-fold cross-validation using the training samples. Considering that the training data are randomly selected, each experiment is repeated 10 times and the average error rates of the two tasks are reported in Fig. 5. The horizontal axis shows the ratio of training samples that are randomly selected to train the metric, and the vertical axis indicates the average test error rate of all tasks.

From the results, we have the following observations:

1. When only 10 % training samples are used, single task learning is incapable of learning a reliable metric and tends to be over-fitting. Its performance is even worse than the Euclidean distance. When more than 20 % training samples are used, its performance is better than Euclidean.
2. Multi-task methods improves the performance especially when the training samples are insufficient. In these experiments, all the multi-task methods demonstrate lower error rates on the test samples than single-task learning.

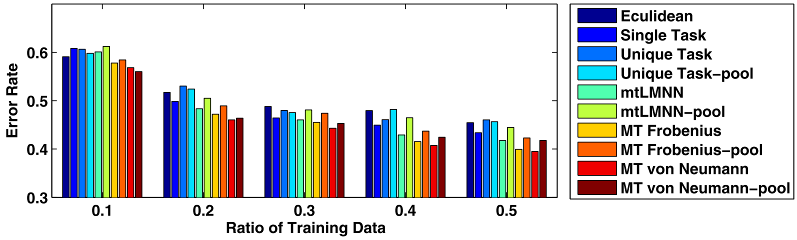


Fig. 5 Test results on multi-speaker vowel classification

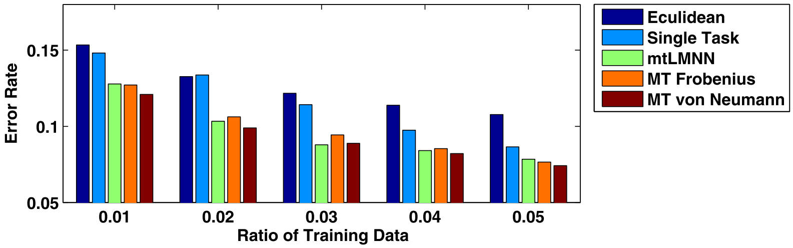


Fig. 6 Test results on handwritten letter classification

3. The pooling version of testing performs better when the samples are very few. For example, MT von Neumann-pool has lower error rate than MT von Neumann when only 10 % samples are used. The more training samples, the worse its performance becomes compared to the no-pooling version. This phenomenon shows that men and women pronounce in an essentially different way and thus mixing them usually deteriorates the classification accuracy.
4. The MT Frobenius approach usually performs better than mtLMNN, which is probably due to the too strict constraint $A_t \succeq B$ as we have explained.
5. The MT von Neumann approach performs the best (including the pooling version) due to its capability to propagate the information among tasks properly.

6.2.2 Handwritten letter classification

Handwritten Letter Classification data set[11] was collected by Rob Kassel at MIT Spoken Language System Group. It consists of 8 binary handwritten letter classification problems where the corresponding letters for each task are: c/e, g/y, m/n, a/g, i/j, a/o, f/t, h/n. The features are the bitmap of the image of written letters and each classification problem is regarded as one task.

The binary labels in different tasks represent different letters and thus this is a label-incompatible problem. Since there is no split training set and test set, we randomly select a proportion of samples to train a metric and use the remaining for test. Because such a split is different for each evaluation, we firstly repeat the experiment 10 times and select the optimal hyper-parameters for each method. Then the hyper-parameters are fixed and the evaluation is repeat 10 times again. The results on the newly split samples are reported in Fig. 6.

[11] http://multitask.cs.berkeley.edu/.

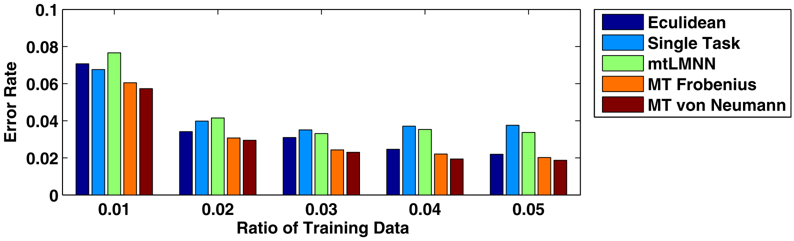


Fig. 7 Test results on USPS digit classification

On this data set, our algorithm performs the best only except when 3 % training samples are used. Even in this case, its accurate is very close to the best. We also observe that the single task method produces high error rate when rather few training samples are used. However, in this dataset, the results on the mtLMNN and the MT Frobenius are very similar. It is possible that the constraint $A_t \succeq B$ is satisfied for this data and thus mtLMNN is more suitable than MT Frobenius. However, both of them perform worse than our method.

6.2.3 USPS digit classification

USPS digit data set[12] consists of 7,291 samples of digits $0 \sim 9$, each of which is a 16×16 grayscale image. Following Zhang and Yeung (2010b), we construct 5 binary classification problems to separate the digits 0/1, 2/3, 4/5, 6/7, 8/9 respectively. Then we learn a metric for each of them jointly. Since each task is to separate a different pair of digits, it is a label-incompatible problem. The experiment setting is as same as Handwritten letter classification in Sect. 6.2.2 and the results are shown in Fig. 7.

For the USPS data set, single-task learning gives very bad performance. It is even worse than the Euclidean metric. This may be due to the over-fitting and thus multi-task learning is needed. The mtLMNN also gives high error rate on this data set, which is sometimes even worse than single-task learning. Since the MT Frobenius exploits the same regularization as mtLMNN but gives a much higher accuracy, it could be caused by the reason that the constraint $A_t \succeq B$ is not satisfied in this data. At last, our method also leads to the best results on all the tests, which again proves its advantage.

6.2.4 Insurance Company Benchmark data

The Insurance Company Benchmark (COIL 2000) data set[13] used in the CoIL 2000 Challenge contains information on customers of an insurance company. The data were collected to predict what kind of people would be interested in buying a caravan insurance policy and consists of 86 variables, including product usage data and socio-demographic data. This dataset consists of 5,822 training samples and 4,000 test samples.

Since each variable is discrete and can be regarded as a label to predict, we consider the problem to predict some of the variables using others as features (Parameswaran and Weinberger 2010). To be specific, we select out a set of related variables from the 86 variables as the targets to predict, and use the other variables as features to predict the selected targets. Prediction of each selected variable is regarded as one task and they constitute a multi-task

[12] http://www-i6.informatik.rwth-aachen.de/~keysers/usps.html.

[13] http://kdd.ics.uci.edu/databases/tic/tic.html.

Table 2 Description of 5 sets of selected targets on CoIL data set

Selected variables	Description
14–15	Variables 14 and 15 describe the data about household without children and with children respectively
16–18	Variables 16, 17, 18 are data about education respect to high level, medium level, and lower level respectively
32–34	Variables 32, 33, 34 are MAUT data about user with 1 car, 2 cars, and no car, respectively
35–36	Variables 35 and 36 are MZ data with respect to national health service and private health insurance respectively
73–74	Variables 73, 74 are number of tractor policies and number of agricultural machines policies respectively

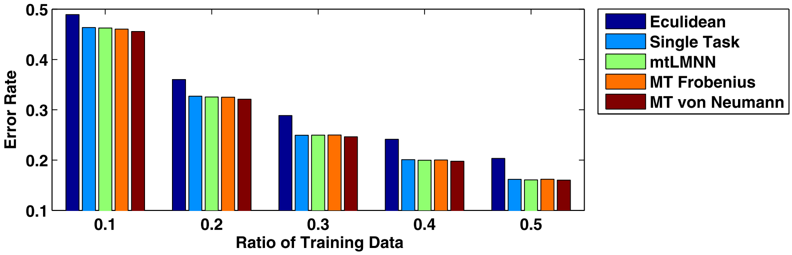


Fig. 8 Results using variable 14–15 as targets

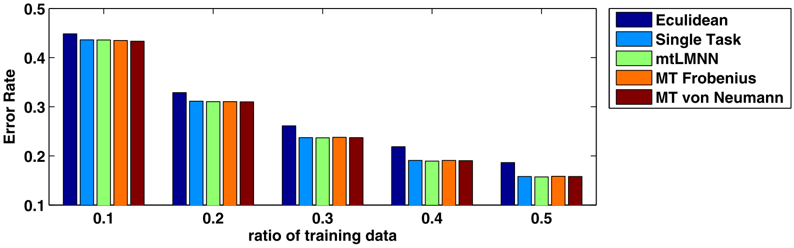


Fig. 9 Results using variable 16–18 as targets

learning problem. Apparently, it is a label-incompatible problem due to the different label sets. This data set has a specified training and test set. We randomly select a certain portion (10 %, 20 %, 30 %, 40 %, 50 %) of samples from the training set to learn a metric and predict the labels of test samples. We construct 5 multi-task learning problems by selecting 5 different target sets of related variables, which can be seen in Table 2. Each experiment is repeated for 10 times and the average error rates are shown in Figs. 8–12.

In these experiment, we can observe that for target sets 14–15, 16–18, 32–24, and 35–36, the test accuracies of all the methods increase with the training samples used. This shows the efficiency of utilizing more training samples. When only 10 % training samples are used, single-task learning does not give an ideal result due to the lack of information. However, the multi-task metric learning methods usually decrease the error rates a bit because it utilizes the information from other tasks. In most of these experiments, our method gives a

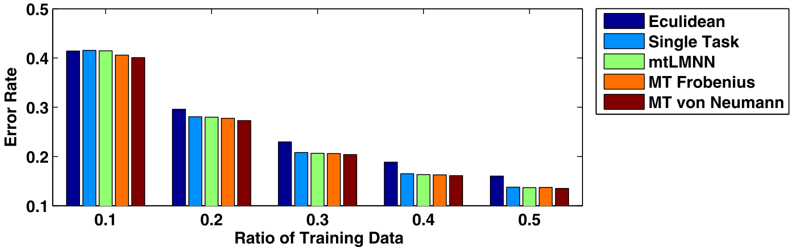


Fig. 10 Results using variable 32–34 as targets

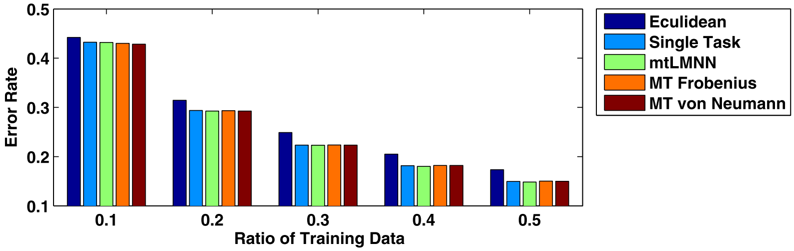


Fig. 11 Results using variable 35–36 as targets

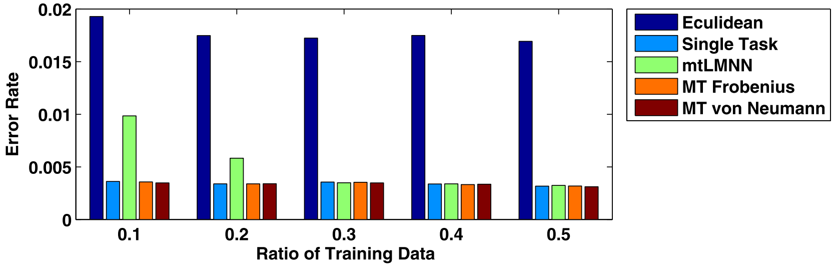


Fig. 12 Results using variable 73–74 as targets

better result than others, but the improvement is very limited. The reason may be that the correlation among these targets are weak.

For the target set 73–74, the results are very different from the former 4 target sets. In this case, all metric learning methods give significantly better performances than Euclidean distance and this proves the advantage of metric learning. However, the multi-task learning methods do not improve result of single-task learning. We try to explain this phenomenon as three possible reasons. (1) Noting that the error rates are almost the same with different number of training samples, information from more training samples cannot improve the performance. Therefore, multi-task learning cannot benefit from propagating more information from other tasks and the accuracy does not increase using multi-task methods. (2) These methods cannot propagate the information among tasks properly. (3) The selected targets are not correlated with others at all. However, we see that even the problem is not appropriate for the multi-task learning, our method doesn't deteriorate the performance of the single-task method.

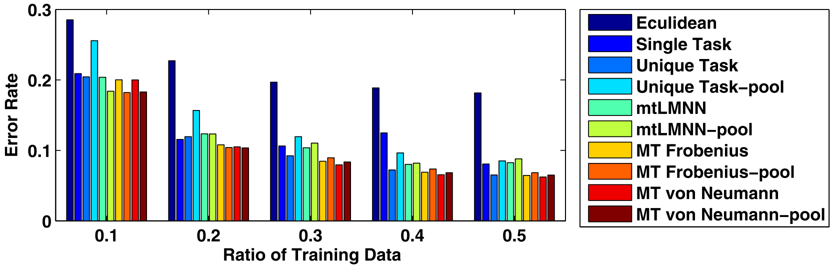


Fig. 13 Test results on Isolet spoken alphabet recognition

6.2.5 Isolet spoken alphabet recognition

In the Isolet data set,[14] 150 speakers spoke the name of each letter of alphabet twice. The task is to classify the letters to be uttered. Since the speakers are grouped into 5 groups, the problem is naturally suitable for multi-task learning where each group is treated as a task. We directly use the data from the website,[15] which has been preprocessed with PCA and split into the training set, validation set, and test set randomly (Parameswaran and Weinberger 2010). To determine the hyper-parameters, we select a specific proportion of training samples to learn a metric with various combinations of hyper-parameters and then test on the validation set. Such an evaluation is repeated 5 times and the hyper-parameters producing the lowest average error rate are chosen as the optimal hyper-parameters. Then the metrics are learned using different proportions of training samples and used for predicting the labels of the test samples. Each experiment is repeated 10 times and the average error rates are shown in Fig. 13.

For this data set, we observe that the unique-task usually generates better performance than single-task, which shows the tasks may be very similar to each other. However, multitask learning can furthermore improve their accuracies. Moreover, MT Frobenius performs better than mtLMNN and MT von Neumann performs even better than MT Frobenius. Our methods (MT von Neumann and MT von Neumann-pool) again demonstrate the best results in most cases. Finally, we found that the pooling version of methods lead to better results when training samples are fewer, which also indicates these tasks are very similar. This implies that, when we have more training samples, it is better to utilize only the samples in this task as prototypes.

7 Conclusion

In this paper, we propose a novel multi-task metric learning framework using Bregman matrix divergence. On one hand, the novel regularized approach extends previous methods from the vector regularization to a general matrix regularization framework; on the other hand and more importantly, by exploiting von Neumann divergence as the regularization, the new multi-task metric learning has the capability to well preserve the data geometry. This leads to more appropriate propagation of side-information among tasks and proves very important for further improving the performance. We propose the concept of *geometry preserving*

[14] Available from UCI Machine Learning Repository.

[15] http://www.cse.wustl.edu/~kilian/code/code.html.

probability (*PG*) and justify our framework with a series of theoretical analysis. Furthermore, our formulation is jointly convex and the global optimal solution can be guaranteed. A series of experiments verify that our proposed approach can significantly outperform the current methods.

Acknowledgements This work has been supported in part by the National Basic Research Program of China (973 Program) Grant 2012CB316301, the National Natural Science Foundation of China (NSFC) Grants 61075052 and 60825301, and Tsinghua National Laboratory for Information Science and Technology (TNList) Cross-discipline Foundation.

Appendix

A.1 Proof of Theorem 3

For convenience of calculating PG, we first define the *geometry preserving indicator*. In the following discussion, we always denote $\mathbf{z}_i = \mathbf{x}_i - \mathbf{y}_i$ as the difference of two points.

Definition 4 (Geometry preserving indicator) The *geometry preserving indicator* $\Psi_{A,B}(\mathbf{x}_1 - \mathbf{y}_1, \mathbf{x}_2 - \mathbf{y}_2) = \Psi_{A,B}(\mathbf{z}_1, \mathbf{z}_2)$ is a function that takes two metrics d_A, d_B as parameters and the differences of two pairs of points as variables. We use d_A and d_B to measure the distances of the two pairs of points and compare which pair of points are relatively further to each other. Then $\Psi = 1$ if the two metrics give the *same* judgement and $\Psi = 0$ *otherwise*. Mathematically, it is defined as

$$\begin{aligned}\Psi_{A,B}(\mathbf{z}_1, \mathbf{z}_2) = \mathbb{1}\Big[&\big(d_A(\mathbf{z}_1) > d_A(\mathbf{z}_2) \wedge d_B(\mathbf{z}_1) > d_B(\mathbf{z}_2)\big)\\ &\vee \big(d_A(\mathbf{z}_1) < d_A(\mathbf{z}_2) \wedge d_B(\mathbf{z}_1) < d_B(\mathbf{z}_2)\big)\\ &\vee \big(d_A(\mathbf{z}_1) = d_A(\mathbf{z}_2) \wedge d_B(\mathbf{z}_1) = d_B(\mathbf{z}_2)\big)\Big],\end{aligned} \tag{19}$$

where $\wedge/\vee$ represents the logical "*and/or*" operator and $\mathbb{1}[\mathcal{E}]$ is the *indicator function* so that

$$\mathbb{1}[\mathcal{E}] = \begin{cases} 1, & \text{if expression } \mathcal{E} \text{ holds;} \\ 0, & \text{if expression } \mathcal{E} \text{ does not hold.} \end{cases}$$

Noting that any $f(\mathbf{x}_1, \mathbf{y}_1, \mathbf{x}_2, \mathbf{y}_2)$ uniquely determines a probability density for their differences with

$$\begin{aligned}\tilde{f}(\mathbf{z}_1, \mathbf{z}_2) &\triangleq \mathrm{P}[\mathbf{x}_1 - \mathbf{y}_1 = \mathbf{z}_1, \mathbf{x}_2 - \mathbf{y}_2 = \mathbf{z}_2]\\ &= \int_{\mathbb{R}^m}\int_{\mathbb{R}^m} f(\mathbf{y}_1 + \mathbf{z}_1, \mathbf{y}_1, \mathbf{y}_2 + \mathbf{z}_2, \mathbf{y}_2)\mathrm{d}\mathbf{y}_1\mathrm{d}\mathbf{y}_2,\end{aligned}$$

the geometry preserving probability $\mathrm{PG}_f(d_A, d_B)$ can be calculated by an integration

$$\mathrm{PG}_f(d_A, d_B) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \Psi_{A,B}(\mathbf{z}_1, \mathbf{z}_2)\tilde{f}(\mathbf{z}_1, \mathbf{z}_2)\mathrm{d}z_1^{(1)}\cdots\mathrm{d}z_1^{(m)}\mathrm{d}z_2^{(1)}\cdots\mathrm{d}z_2^{(m)}, \tag{20}$$

where $\mathbf{z}_i = [z_i^{(1)}\ z_i^{(2)}\ \ldots\ z_i^{(m)}]^\top$ and the superscript indicates the dimension of the coordinate.

Then we can give the proof of Theorem 3.

Proof of Theorem 3 For the integration to calculate PG shown in (20), the Cartesian coordinate system can be transformed to the spherical coordinate system by the following transformation (Wikipedia 2013a; Kingravi 2007):

$$\begin{aligned}
z_i^{(1)} &= r_i q_i^{(1)} = r_i \cos\bigl(\psi_i^{(1)}\bigr)\\
z_i^{(2)} &= r_i q_i^{(2)} = r_i \sin\bigl(\psi_i^{(1)}\bigr)\cos\bigl(\psi_i^{(2)}\bigr)\\
&\vdots\\
z_i^{(k)} &= r_i q_i^{(k)} = r_i \left(\prod_{j=1}^{k-1} \sin\bigl(\psi_i^{(j)}\bigr)\right)\cos\bigl(\psi_i^{(k)}\bigr)\\
&\vdots\\
z_i^{(m-1)} &= r_i q_i^{(m-1)} = r_i \sin\bigl(\psi_i^{(1)}\bigr)\sin\bigl(\psi_i^{(2)}\bigr)\cdots\sin\bigl(\psi_i^{(d-2)}\bigr)\cos\bigl(\psi_i^{(d-1)}\bigr)\\
z_i^{(m)} &= r_i q_i^{(m)} = r_i \sin\bigl(\psi_i^{(1)}\bigr)\sin\bigl(\psi_i^{(2)}\bigr)\cdots\sin\bigl(\psi_i^{(d-2)}\bigr)\sin\bigl(\psi_i^{(d-1)}\bigr)
\end{aligned}$$

Then we have $\mathbf{z}_i = r_i\mathbf{q}_i$ ($r_i \geq 0$) and $\|\mathbf{q}_i\| = 1$ which means r_i is the length of $\mathbf{z}_i$ and $\mathbf{q}_i$ is a unit vector indicating the direction of $\mathbf{z}_i$. The volume element can be calculated by a Jacobian determinant (Kingravi 2007) as

$$r_i^{m-1}\sin^{m-2}\bigl(\psi_i^{(1)}\bigr)\cdots\sin\bigl(\psi_i^{(m-2)}\bigr)\mathrm{d}r_i\,\mathrm{d}\psi_i^{(1)}\cdots\mathrm{d}\psi_i^{(m-1)} \triangleq r_i^{m-1}\mathrm{d}r_i\,\mathrm{d}\Omega(\mathbf{q}_i),$$

where $\mathrm{d}\Omega(\mathbf{q}_i) = \sin^{m-2}(\psi_i^{(1)})\cdots\sin(\psi_i^{(m-2)})\mathrm{d}\psi_i^{(1)}\cdots\mathrm{d}\psi_i^{(m-1)}$ is the *solid angle* element corresponding to the direction determined by $\{\psi_i\}_{i=1}^{m-1}$, or equivalently by $\mathbf{q}_i$.

Then the geometry preserving probability (20) can be reformulated by coordinate transformation as

$$\begin{aligned}
&\mathrm{PG}_f(d_A, d_B)\\
&\quad= \int_{\psi_2^{(m-1)}=0}^{2\pi}\int_{\psi_2^{(m-2)}=0}^{\pi}\cdots\int_{\psi_2^{(1)}=0}^{\pi}\int_{\psi_1^{(m-1)}=0}^{2\pi}\int_{\psi_1^{(m-2)}=0}^{\pi}\cdots\int_{\psi_1^{(1)}=0}^{\pi}\\
&\qquad\times\int_{r_2=0}^{\infty}\int_{r_1=0}^{\infty}\Psi_{A,B}(r_1\mathbf{q}_1, r_2\mathbf{q}_2)\tilde{f}(r_1\mathbf{q}_1, r_2\mathbf{q}_2)\\
&\qquad\cdot r_1^{m-1}\sin^{m-2}\bigl(\psi_1^{(1)}\bigr)\cdots\sin\bigl(\psi_1^{(m-2)}\bigr)r_2^{m-1}\sin^{m-2}\bigl(\psi_2^{(1)}\bigr)\cdots\sin\bigl(\psi_2^{(m-2)}\bigr)\\
&\qquad\cdot\mathrm{d}r_1\mathrm{d}r_2\mathrm{d}\psi_1^{(1)}\cdots\mathrm{d}\psi_1^{(m-2)}\mathrm{d}\psi_1^{(m-1)}\psi_2^{(1)}\cdots\mathrm{d}\psi_2^{(m-2)}\mathrm{d}\psi_2^{(m-1)}\\
&\quad= \iint_{\mathbb{S}^{m-1}\times\mathbb{S}^{m-1}}\iint_{\mathbb{R}_+\times\mathbb{R}_+}\Psi_{A,B}(r_1\mathbf{q}_1, r_2\mathbf{q}_2)\tilde{f}(r_1\mathbf{q}_1, r_2\mathbf{q}_2)r_1^{m-1}r_2^{m-1}\mathrm{d}r_1\mathrm{d}r_2\mathrm{d}\Omega(\mathbf{q}_1)\mathrm{d}\Omega(\mathbf{q}_2)\\
&\quad\triangleq \iint_{\mathbb{S}^{m-1}\times\mathbb{S}^{m-1}} R_{\mathbf{q}_1,\mathbf{q}_2}(A, B)\mathrm{d}\Omega(\mathbf{q}_1)\mathrm{d}\Omega(\mathbf{q}_2),
\end{aligned}$$

where

$$R_{\mathbf{q}_1,\mathbf{q}_2}(A, B) = \iint_{\mathbb{R}_+\times\mathbb{R}_+}\Psi_{A,B}(r_1\mathbf{q}_1, r_2\mathbf{q}_2)\tilde{f}(r_1\mathbf{q}_1, r_2\mathbf{q}_2)r_1^{m-1}r_2^{m-1}\mathrm{d}r_1\mathrm{d}r_2,$$

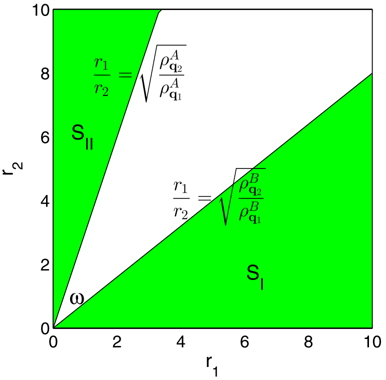


Fig. 14 Domain of integration of $\mathbf{r}$ in $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$

and the field of integration of $\mathbf{q}_i$ is $\mathbb{S}^{m-1}$ is an $(m-1)$-dimensional unit sphere.

By integrating the radial variables r_1, r_2 firstly as the equation above, $\mathrm{PG}_f(d_A, d_B)$ is reformulated as an integration of $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ in different pairs of directions. For $\forall i = 1, 2$, we have $d_A^2(r_i\mathbf{q}_i) = r_i^2\mathbf{q}_i^\top A\mathbf{q}_i = r_i^2\rho^A_{\mathbf{q}_i}$, and thus the squared distance $d_A^2(r_i\mathbf{q}_i)$ equals to the weighted scale on $\mathbf{q}_i$ with weight r_i^2. Similarly, $d_B^2(r_i\mathbf{q}_i) = r_i^2\rho^B_{\mathbf{q}_i}$. It is straightforward to show that

$$d_A(r_1\mathbf{q}_1) > d_A(r_2\mathbf{q}_2) \quad\Leftrightarrow\quad \frac{r_1}{r_2} > \sqrt{\frac{\rho^A_{\mathbf{q}_2}}{\rho^A_{\mathbf{q}_1}}}, \qquad d_B(r_1\mathbf{q}_1) > d_B(r_2\mathbf{q}_2) \quad\Leftrightarrow\quad \frac{r_1}{r_2} > \sqrt{\frac{\rho^B_{\mathbf{q}_2}}{\rho^B_{\mathbf{q}_1}}}.$$

Therefore, denoting $\mathbf{r} = [r_1\ r_2]^\top$, the geometry preserving indicator is

$$\Psi_{A,B}(r_1\mathbf{q}_1, r_2\mathbf{q}_2) = \begin{cases} 1, & \text{if } \mathbf{r} \in S_{\mathrm{I}} \cup S_{\mathrm{II}}; \\ 0, & \text{otherwise,} \end{cases}$$

where

$$S_{\mathrm{I}} = \left\{\mathbf{r}\left|\frac{r_1}{r_2} > \max\left\{\sqrt{\frac{\rho^A_{\mathbf{q}_2}}{\rho^A_{\mathbf{q}_1}}}, \sqrt{\frac{\rho^B_{\mathbf{q}_2}}{\rho^B_{\mathbf{q}_1}}}\right\}\right.\right\}, \qquad S_{\mathrm{II}} = \left\{\mathbf{r}\left|\frac{r_1}{r_2} < \min\left\{\sqrt{\frac{\rho^A_{\mathbf{q}_2}}{\rho^A_{\mathbf{q}_1}}}, \sqrt{\frac{\rho^B_{\mathbf{q}_2}}{\rho^B_{\mathbf{q}_1}}}\right\}\right.\right\}.$$

Since $\mathbf{q}_1, \mathbf{q}_2$ are fixed here, $\tilde{f}$ is a function of $\mathbf{r}$. Then $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ can be reformulated as

$$R_{\mathbf{q}_1,\mathbf{q}_2}(A,B) = \int_{S_{\mathrm{I}}\cup S_{\mathrm{II}}} \tilde{f}(\mathbf{r}) r_1^{m-1} r_2^{m-1} \mathrm{d}\mathbf{r}. \tag{21}$$

The domain of integration is illustrated as the shadow region in Fig. 14, which is determined by two boundaries corresponding to d_A and d_B respectively. If d_B (or d_A) is given, the corresponding boundary is fixed, then considering that the integral function $\tilde{f}(\mathbf{r})r_1^{m-1}r_2^{m-1}$ is non-negative anywhere, $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ monotonically decreases with $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$, which is the angle between the two boundaries determined by $\rho^A_{\mathbf{q}_i}$ and $\rho^B_{\mathbf{q}_i}$ as

$$\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B) = \arctan\sqrt{\frac{\rho^A_{\mathbf{q}_2}}{\rho^A_{\mathbf{q}_1}}} - \arctan\sqrt{\frac{\rho^B_{\mathbf{q}_2}}{\rho^B_{\mathbf{q}_1}}}.$$

Furthermore, when d_B (or d_A) is given, $\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ uniquely determines the integral domain S_{I} and S_{II}, and thus

$$\begin{aligned}
\left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A_1,B)\right| > \left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A_2,B)\right| &\Rightarrow R_{\mathbf{q}_1,\mathbf{q}_2}(A_1,B) \le R_{\mathbf{q}_1,\mathbf{q}_2}(A_2,B),\\
\left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A_1,B)\right| < \left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A_2,B)\right| &\Rightarrow R_{\mathbf{q}_1,\mathbf{q}_2}(A_1,B) \ge R_{\mathbf{q}_1,\mathbf{q}_2}(A_2,B),\\
\left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B_1)\right| > \left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B_2)\right| &\Rightarrow R_{\mathbf{q}_1,\mathbf{q}_2}(A,B_1) \le R_{\mathbf{q}_1,\mathbf{q}_2}(A,B_2),\\
\left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B_1)\right| < \left|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B_2)\right| &\Rightarrow R_{\mathbf{q}_1,\mathbf{q}_2}(A,B_1) \ge R_{\mathbf{q}_1,\mathbf{q}_2}(A,B_2),
\end{aligned}$$

which indicates that $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ *monotonically decreases* with $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$ and the proof completes. □

Remark Note that when d_A and d_B are both unknown and learned simultaneously, a smaller angle $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$ does not guarantee a greater $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ strictly, which also depends on f. However, without precise information about f, a smaller $|\omega_{\mathbf{q}_1,\mathbf{q}_2}(A,B)|$ usually accompanies with a greater $R_{\mathbf{q}_1,\mathbf{q}_2}(A,B)$ in general.

A.2 Proof of Theorem 4

Although (11) and (12) in Theorem 4 can be unified as (10), the proofs of them are quite different and we will present them separately.

First, (11) can be proved using some simple matrix calculations.

Proof of Inequality (11) Assuming the spectral decomposition of $A-B$ is $U\Lambda U^\top$ and $\tilde{\mathbf{w}}_i = U^\top \mathbf{w}_i$, for any unit vector $\mathbf{w}_i$, we have

$$\begin{aligned}
\mathbf{w}_i^\top A\mathbf{w}_i - \mathbf{w}_i^\top B\mathbf{w}_i &= \mathbf{w}_i^\top (A-B)\mathbf{w}_i\\
&= \mathbf{w}_i^\top U\Lambda U^\top \mathbf{w}_i\\
&= \left(U^\top \mathbf{w}_i\right)^\top \Lambda \left(U^\top \mathbf{w}_i\right)\\
&= \tilde{\mathbf{w}}_i^\top \Lambda \tilde{\mathbf{w}}_i\\
&= \sum_{j=1}^m \tilde{w}_{ij}^2 \lambda_j .
\end{aligned}$$

For any group of orthonormal basis W, it is straightforward that $\tilde{W} = [\tilde{\mathbf{w}}_1 \ \dots \ \tilde{\mathbf{w}}_m] = U^\top W$ is also orthonormal, and thus the element-wise squared matrix of $\tilde{W}$ is a *double stochastic matrix* (Marshall et al. 2011) satisfying $\sum_{i=1}^m \tilde{w}_{ij}^2 = 1, \forall j$ and $\sum_{j=1}^m \tilde{w}_{ij}^2 = 1, \forall i$.

Therefore, using Jensen's inequality (Hardy et al. 1988) and by the convexity of $f(x) = x^2$, we obtain

$$\begin{aligned}
\left\|\rho_W^A - \rho_W^B\right\|_2^2 &= \sum_{i=1}^m \left(\mathbf{w}_i^\top A\mathbf{w}_i - \mathbf{w}_i^\top B\mathbf{w}_i\right)^2\\
&= \sum_{i=1}^m \left(\sum_{j=1}^m \tilde{w}_{ij}^2 \lambda_j\right)^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i=1}^{m}\sum_{j=1}^{m} \tilde{w}_{ij}^2 \lambda_j^2 \\
&= \sum_{j=1}^{m}\sum_{i=1}^{m} \tilde{w}_{ij}^2 \lambda_j^2 \\
&= \sum_{j=1}^{m} \lambda_j^2 \\
&= \|A - B\|_{\mathrm{F}}^2.
\end{aligned}
$$

The equality holds if and only if W is composed of the eigenvectors of $A - B$, which is straightforward by substituting $W = [\mathbf{u}_1 \ \ldots \ \mathbf{u}_m]$ into the formula above. Thus $\|A - B\|_{\mathrm{F}}^2$ provides a strict upper bound for $\|\rho_W^A - \rho_W^B\|^2$ and the proof completes. □

Then, we focus on the proof of (12), which is supported by the following lemma.

Lemma 5 *For any* trace preserving map $\varPhi$ *defined as* $\varPhi(A) = \sum_{i=1}^n V_i A V_i^\top$ *where* $\sum_{i=1}^n V_i^\top V_i = \mathbf{I}_m$ (Lindblad 1975) *and* A, B *are both SPD matrices, we have that* $D_{\mathrm{vN}}(\varPhi(A), \varPhi(B)) \leq D_{\mathrm{vN}}(A, B)$.

Indeed, a very similar result as Lemma 5 has been intensively studied in quantum information theory (Nielsen and Chuang 2010; Lindblad 1975), which reveals the relationship between the von Neumann divergence (also referred as *quantum relative entropy*) of two *quantum densities* (SPD matrices with trace 1) and the KL-divergence (also referred as *relative entropy*) of their *measurements* (vectors). It's notable that there exist two differences compared with (12):

1. The quantum density is represented as a SPD matrix with trace 1 while the trace of a Mahalanobis matrix we studied could be any non-negative real number.
2. The von Neumann divergence in quantum information theory is defined as $\bar{D}_{\mathrm{vN}}(A, B) = \mathrm{tr}(A \log A - A \log B)$. Although it is equivalent to $D_{\mathrm{vN}}(A, B)$ for quantum densities satisfying $\mathrm{tr}(A) = \mathrm{tr}(B) = 1$, they are different for Mahalanobis matrices in general.

Due to these differences, the conclusion in Lindblad (1975) cannot be used directly.

Following a similar way of Lindblad (1975), we give the proof of Lemma 5 after a series of necessary definitions and lemmas.

Lemma 6 (Logarithm of Kronecker product) *Suppose* $A_{m\times m}$ *and* $B_{n\times n}$ *are both symmetric positive definite matrices*, *then the matrix logarithm of their Kronecker product* (Horn and Johnson 1991) $A \otimes B$ *has the following decomposition*.

$$\log(A \otimes B) = \log A \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \log B. \tag{22}$$

Proof Assume that the spectral decomposition of A and B are $A = V\Lambda V^\top$ and $B = U\Theta U^\top$ respectively, then

$$\begin{aligned}
\log(A \otimes B) &= \log\big(\big(V\Lambda V^\top\big) \otimes \big(U\Theta U^\top\big)\big)\\
&= \log\big((V \otimes U)(\Lambda \otimes \Theta)(V \otimes U)^\top\big)\\
&= (V \otimes U)\log\big(\operatorname{diag}[\lambda_1, \dots, \lambda_m] \otimes \operatorname{diag}[\theta_1, \dots, \theta_n]\big)\big(V^\top \otimes U^\top\big)\\
&= (V \otimes U)\log\big(\operatorname{diag}[\lambda_1\theta_1, \dots, \lambda_1\theta_n, \dots, \lambda_m\theta_1, \dots, \lambda_m\theta_n]\big)\big(V^\top \otimes U^\top\big)\\
&= (V \otimes U)\operatorname{diag}[\log\lambda_1 + \log\theta_1, \dots, \log\lambda_m + \log\theta_n]\big(V^\top \otimes U^\top\big)\\
&= (V \otimes U)(\log\Lambda \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \log\Theta)\big(V^\top \otimes U^\top\big)\\
&= \big(V\log\Lambda V^\top\big) \otimes \big(U\mathbf{I}_n U^\top\big) + \big(V\mathbf{I}_m V^\top\big) \otimes \big(U\log\Theta U^\top\big)\\
&= \log A \otimes \mathbf{I}_n + \mathbf{I}_m \otimes \log B,
\end{aligned}$$

and this completes the proof. □

Lemma 7 (Additivity of von Neumann divergence) *Suppose that A_1, A_2, B_1, B_2 are $m \times m$ SPD matrices, then the following equation holds.*

$$\begin{aligned}
D_{\mathrm{vN}}(A_1 \otimes A_2, B_1 \otimes B_2) = {} & D_{\mathrm{vN}}(A_1, B_1) \cdot \operatorname{tr}A_2 + D_{\mathrm{vN}}(A_2, B_2) \cdot \operatorname{tr}A_1\\
& + (\operatorname{tr}A_1 - \operatorname{tr}B_1)(\operatorname{tr}A_2 - \operatorname{tr}B_2).
\end{aligned}$$

Specifically, if $A_2 = B_2 = P$, the equation is simplified as

$$D_{\mathrm{vN}}(A_1 \otimes P, B_1 \otimes P) = D_{\mathrm{vN}}(A_1, B_1) \cdot \operatorname{tr}P.$$

Proof Using (22), we have

$$\begin{aligned}
& D_{\mathrm{vN}}(A_1 \otimes A_2, B_1 \otimes B_2)\\
&\quad = \operatorname{tr}\big((A_1 \otimes A_2)\big(\log(A_1 \otimes A_2) - \log(B_1 \otimes B_2)\big) - A_1 \otimes A_2 + B_1 \otimes B_2\big)\\
&\quad = \operatorname{tr}\big((A_1 \otimes A_2)\big((\log A_1 - \log B_1) \otimes \mathbf{I}_m + \mathbf{I}_m \otimes (\log A_2 - \log B_2)\big)\\
&\qquad - A_1 \otimes A_2 + B_1 \otimes B_2\big)\\
&\quad = \operatorname{tr}\big(A_1(\log A_1 - \log B_1) \otimes A_2 + A_1 \otimes A_2(\log A_2 - \log B_2) - A_1 \otimes A_2 + B_1 \otimes B_2\big)\\
&\quad = D_{\mathrm{vN}}(A_1, B_1) \cdot \operatorname{tr}A_2 + D_{\mathrm{vN}}(A_2, B_2) \cdot \operatorname{tr}A_1\\
&\qquad + \operatorname{tr}(A_1 \otimes A_2 - B_1 \otimes A_2 - A_1 \otimes B_2 + B_1 \otimes B_2)\\
&\quad = D_{\mathrm{vN}}(A_1, B_1) \cdot \operatorname{tr}A_2 + D_{\mathrm{vN}}(A_2, B_2) \cdot \operatorname{tr}A_1 + (\operatorname{tr}A_1 - \operatorname{tr}B_1)(\operatorname{tr}A_2 - \operatorname{tr}B_2).
\end{aligned}$$

If $A_2 = B_2 = P$, then

$$D_{\mathrm{vN}}(A_1 \otimes P, B_1 \otimes P) = D_{\mathrm{vN}}(A_1, B_1) \cdot \operatorname{tr}P$$

is straightforward. □

Lemma 8 (Unitary operation invariant property) *The von Neumann divergence is invariant under unitary operation.*

To be specific, assume that $U \in \mathbb{C}^{n\times n}$ is a unitary matrix satisfying $UU^{\mathrm{H}} = U^{\mathrm{H}}U = \mathbf{I}_n$, where U^{H} is the conjugate transpose (*also called* associate matrix *in quantum*) *of U. Then, the following formula holds for any symmetric positive definite matrices $A, B \in \mathrm{Pos}(m)$.*

$$D_{\mathrm{vN}}\big(UAU^{\mathrm{H}}, UBU^{\mathrm{H}}\big) = D_{\mathrm{vN}}(A, B).$$

Proof It is straightforward to verify that $\log(UAU^{\mathrm{H}}) = U \log AU^{\mathrm{H}}$ and thus

$$\begin{aligned}
D_{\mathrm{vN}}\big(UAU^{\mathrm{H}}, UBU^{\mathrm{H}}\big) &= \mathrm{tr}\big(UAU^{\mathrm{H}}\log\big(UAU^{\mathrm{H}}\big) - UAU^{\mathrm{H}}\log\big(UBU^{\mathrm{H}}\big) \\
&\quad - UAU^{\mathrm{H}} + UBU^{\mathrm{H}}\big) \\
&= \mathrm{tr}\big(UAU^{\mathrm{H}}U\log AU^{\mathrm{H}} - UAU^{\mathrm{H}}U\log BU^{\mathrm{H}} - UAU^{\mathrm{H}} + UBU^{\mathrm{H}}\big) \\
&= \mathrm{tr}\big(U(A\log A - A\log B - A + B)U^{\mathrm{H}}\big) \\
&= \mathrm{tr}\big((A\log A - A\log B - A + B)\big(U^{\mathrm{H}}U\big)\big) \\
&= \mathrm{tr}(A\log A - A\log B - A + B) \\
&= D_{\mathrm{vN}}(A, B).
\end{aligned}$$

Then the invariant property of von Neumann divergence under unitary operation is proved. □

Definition 5 (Partial trace, Example 2.1 of Watrous (2008))[16] Assume that $\mathbb{X} = \mathbb{C}^m$ and $\mathbb{Y} = \mathbb{C}^n$ are complex spaces, then $A \in \mathrm{L}(\mathbb{X}) = \mathbb{C}^{m\times m}$ and $B \in \mathrm{L}(\mathbb{Y}) = \mathbb{C}^{n\times n}$ define a linear transformation in $\mathbb{X}$ and $\mathbb{Y}$ respectively. The *partial trace* on $\mathbb{X}$ is an operator from $\mathrm{L}(\mathbb{X} \otimes \mathbb{Y}) = \mathbb{C}^{mn\times mn}$ to $\mathrm{L}(\mathbb{Y}) = \mathbb{C}^{n\times n}$ that satisfies

$$(\mathrm{tr} \otimes \mathbf{I}_{\mathbb{Y}})(A \otimes B) = \mathrm{tr}(A)B$$

for all $A \in \mathrm{L}(\mathbb{X})$ and $B \in \mathrm{L}(\mathbb{Y})$. It is more commonly denoted $\mathrm{Tr}_{\mathbb{X}}$ and may alternately expressed as follows.

Assuming that $\{\mathbf{u}_i\}_{i=1}^m$ is any orthonormal basis of $\mathbb{X}$, then

$$\mathrm{Tr}_{\mathbb{X}}(A) = \sum_i \big(\mathbf{u}_i^{\mathrm{H}} \otimes \mathbf{I}_{\mathbb{Y}}\big)A\big(\mathbf{u}_i^{\mathrm{H}} \otimes \mathbf{I}_{\mathbb{Y}}\big)$$

for all $A \in \mathrm{L}(\mathbb{X} \otimes \mathbb{Y})$.

Lemma 9 (Monotonicity of von Neumann divergence) *Let $\mathbb{X}$, $\mathbb{Y}$ are two complex Euclidean spaces. For any choice of SPD matrices $A, B \in \mathrm{Pos}(\mathbb{X} \otimes \mathbb{Y})$, we have*

$$D_{\mathrm{vN}}\big(\mathrm{Tr}_{\mathbb{Y}}(A)\big\|\mathrm{Tr}_{\mathbb{Y}}(B)\big) \le D_{\mathrm{vN}}(A, B).$$

Proof Define the completely depolarizing operation $\Omega(A) = \frac{1}{m}\mathrm{tr}(A)\mathbf{I}_{\mathbb{X}}$. Then by Lecture 6 & 9 of Watrous (2008), Ω and $\mathbf{I} \otimes \Omega$ are both *mixed unitary*, which means that there exists a

[16] The original definition is in the language of quantum information and we translate it into the language of matrix theory for convenience.

collection $U_1, \dots, U_N \in U(\mathbb{X})$ of unitary matrices on $\mathbb{X}$ and a probability vector $p_1, \dots, p_N$ such that

$$(\mathbf{I}_{\mathbb{X}} \otimes \Omega)(A) = \sum_{j=1}^{N} p_j U_j A U_j^{\mathrm{H}}.$$

In another aspect, $\mathbf{I}_{\mathbb{X}} \otimes \Omega(A) = \mathrm{Tr}_{\mathbb{Y}}(A) \otimes \frac{\mathbf{I}_{\mathbb{Y}}}{n}$ where $n = \dim(\mathbb{Y})$.

Therefore, we have

$$\begin{aligned} D_{\mathrm{vN}}\big(\mathrm{Tr}_{\mathbb{Y}}(A), \mathrm{Tr}_{\mathbb{Y}}(B)\big) &= D_{\mathrm{vN}}\left(\mathrm{Tr}_{\mathbb{Y}}(A) \otimes \frac{\mathbf{I}_{\mathbb{Y}}}{n}, \mathrm{Tr}_{\mathbb{Y}}(B) \otimes \frac{\mathbf{I}_{\mathbb{Y}}}{n}\right) \Big/ \mathrm{tr}\left(\frac{\mathbf{I}_{\mathbb{Y}}}{n}\right) \\ &= D_{\mathrm{vN}}\big(\mathbf{I}_{\mathbb{X}} \otimes \Omega(A), \mathbf{I}_{\mathbb{X}} \otimes \Omega(B)\big) \\ &= D_{\mathrm{vN}}\left(\sum_{j=1}^{N} p_j U_j A U_j^{\mathrm{H}}, \sum_{j=1}^{N} p_j U_j B U_j^{\mathrm{H}}\right) \\ &\le \sum_{j=1}^{N} p_j D_{\mathrm{vN}}\big(U_j A U_j^{\mathrm{H}}, U_j B U_j^{\mathrm{H}}\big) \\ &= \sum_{j=1}^{N} p_j D_{\mathrm{vN}}(A, B) \\ &= D_{\mathrm{vN}}(A, B), \end{aligned}$$

where the first equality comes from Lemma 7, the inequality is due to the joint convexity of von Neumann divergence (Theorem 1), and the following equality comes from Lemma 8. Then we obtain the conclusion. □

Now we are ready to prove Lemma 5.

Proof of Lemma 5 Assume that $\{\mathbf{r}_i, i = 1, \dots, n\}$ is an orthonormal basis of $\mathbb{R}^n$ and $\mathbf{s} \in \mathbb{R}^n$ is an arbitrary unit vector. Define $Q = \sum_{i=1}^{n} V_i \otimes \mathbf{r}_i \mathbf{s}^{\top}$, then we have

$$Q^{\top} Q = \sum_{ij} V_i^{\top} V_j \otimes \mathbf{s}\mathbf{r}_i^{\top} \mathbf{r}_j \mathbf{s}^{\top} = \left(\sum_i V_i^{\top} V_i\right) \otimes \mathbf{s}\mathbf{s}^{\top} = \big(\mathbf{I}_m \otimes \mathbf{s}\mathbf{s}^{\top}\big)^{\top} \big(\mathbf{I}_m \otimes \mathbf{s}\mathbf{s}^{\top}\big),$$

and thus there is an orthogonal matrix $U \in (\mathbb{R}^m \otimes \mathbb{R}^n) \times (\mathbb{R}^m \otimes \mathbb{R}^n)$ such that $Q = U(\mathbf{I}_m \otimes \mathbf{s}\mathbf{s}^{\top})$.

Consequently,

$$\begin{aligned} U\big(A \otimes \mathbf{s}\mathbf{s}^{\top}\big)U^{\top} &= U\big(\mathbf{I}_m \otimes \mathbf{s}\mathbf{s}^{\top}\big)(A \otimes \mathbf{I}_n)\big(\mathbf{I}_m \otimes \mathbf{s}\mathbf{s}^{\top}\big)U^{\top} = Q(A \otimes \mathbf{I}_n)Q^{\top} \\ &= \sum_{ij} V_i A V_j^{\top} \otimes \mathbf{r}_i \mathbf{s}^{\top} \mathbf{I}_n \mathbf{s} \mathbf{r}_j^{\top} = \sum_{ij} V_i A V_j^{\top} \otimes \mathbf{r}_i \mathbf{r}_j^{\top}, \end{aligned}$$

and then from $\mathrm{tr}(\mathbf{r}_i \mathbf{r}_j^{\top}) = \delta_{ij}$, we have

$$\mathrm{Tr}_{\mathbb{Y}}\big(U\big(A \otimes \mathbf{s}\mathbf{s}^{\top}\big)U^{\top}\big) = \sum_i V_i A V_i^{\top} = \Phi(A).$$

Thus, the KL-divergence can be calculated as

$$\begin{aligned} D_{\mathrm{vN}}\big(\varPhi(A), \varPhi(B)\big) &= D_{\mathrm{vN}}\big(\mathrm{Tr}_{\mathbb{Y}}\big(U\big(A \otimes \mathbf{s}\mathbf{s}^\top\big)U^\top\big), \mathrm{Tr}_{\mathbb{Y}}\big(U\big(B \otimes \mathbf{s}\mathbf{s}^\top\big)U^\top\big)\big) \\ &\le D_{\mathrm{vN}}\big(U\big(A \otimes \mathbf{s}\mathbf{s}^\top\big)U^\top, U\big(B \otimes \mathbf{s}\mathbf{s}^\top\big)U^\top\big) \\ &= D_{\mathrm{vN}}\big(A \otimes \mathbf{s}\mathbf{s}^\top, B \otimes \mathbf{s}\mathbf{s}^\top\big) \\ &= D_{\mathrm{vN}}(A, B) \cdot \mathrm{tr}\big(\mathbf{s}\mathbf{s}^\top\big) \\ &= D_{\mathrm{vN}}(A, B) \end{aligned}$$

and this completes the proof. □

Then, we are ready to prove (12) using the above results.

Proof of Inequality (12) Define the projectors $W_i = \mathbf{w}_i \mathbf{w}_i^\top$, then we have

$$\sum_i W_i^\top W_i = \sum_i \mathbf{w}_i \mathbf{w}_i^\top = W W^\top = \mathbf{I}_m$$

and thus $\varPhi(A) = \sum_{i=1}^m W_i A W_i^\top$ is a trace preserving map.

Then it is straightforward to verify that

$$\begin{aligned} D_{\mathrm{KL}}\big(\rho_W^A, \rho_W^B\big) &= \sum_i D_{\mathrm{KL}}\big(\mathbf{w}_i^\top A \mathbf{w}_i, \mathbf{w}_i^\top B \mathbf{w}_i\big) \\ &= \sum_{i,j} \big(\mathbf{w}_i^\top \mathbf{w}_j\big)^2 D_{\mathrm{KL}}\big(\mathbf{w}_i^\top A \mathbf{w}_i, \mathbf{w}_j^\top B \mathbf{w}_j\big) \\ &= D_{\mathrm{vN}}\Big(\sum_i \mathbf{w}_i \big(\mathbf{w}_i^\top A \mathbf{w}_i\big)\mathbf{w}_i^\top, \sum_j \mathbf{w}_j \big(\mathbf{w}_j^\top B \mathbf{w}_j\big)\mathbf{w}_j^\top\Big) \\ &= D_{\mathrm{vN}}\Big(\sum_i W_i A W_i^\top, \sum_i W_i B W_i^\top\Big) \\ &\le D_{\mathrm{vN}}(A, B), \end{aligned}$$

where the third equality is from (9) and the last inequality results from Lemma 5, and this completes the proof. □

Combining the proofs of (11) and (12), Theorem 4 is proved.

A.3 Proofs of Proposition 1 and Proposition 2

We will first prove Proposition 1 after a prerequisite lemma.

Lemma 10 *Supposing that $f : \mathbb{X} \to \mathbb{R}$, $g : \mathbb{X} \to (0, +\infty)$ are functions on $\mathbb{X}$, $\mathcal{X} \subset \mathbb{X}$ is a subset of $\mathbb{X}$, and $a, b \in \mathbb{R}$ are two real numbers, we have*

$$a < \frac{f(\mathbf{x})}{g(\mathbf{x})} < b, \quad \forall \mathbf{x} \in \mathcal{X} \Rightarrow a < \frac{\int_{\mathcal{X}} f(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x})}{\int_{\mathcal{X}} g(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x})} < b.$$

Proof Since $g(x) > 0$, it is straightforward that

$$\begin{aligned} & ag(\mathbf{x}) < f(\mathbf{x}) < bg(\mathbf{x}), \quad \forall \mathbf{x} \in \mathcal{X} \\ & \Rightarrow \int_{\mathcal{X}} ag(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}) < \int_{\mathcal{X}} f(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}) < \int_{\mathcal{X}} bg(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}). \end{aligned}$$

Then from

$$\int_{\mathcal{X}} ag(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}) = a \int_{\mathcal{X}} g(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}) > 0, \qquad \int_{\mathcal{X}} bg(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}) = b \int_{\mathcal{X}} g(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x}) > 0,$$

we obtain

$$a < \frac{\int_{\mathcal{X}} f(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x})}{\int_{\mathcal{X}} g(\mathbf{x}) \mathrm{d}\sigma(\mathbf{x})} < b,$$

which completes the proof. □

Proof of Proposition 1 The integration element of $\mathcal{E}(A, B)$ with respect to $\rho_{\mathbf{q}}^A$ is

$$\begin{aligned} & \left(\int_{\mathbb{S}^{m-1}} \left| \omega_{\mathbf{q},\mathbf{q}_2}(A, B) \right| \mathrm{d}\Omega(\mathbf{q}_2) \right) \mathrm{d}\Omega(\mathbf{q}) + \left(\int_{\mathbb{S}^{m-1}} \left| \omega_{\mathbf{q}_1,\mathbf{q}}(A, B) \right| \mathrm{d}\Omega(\mathbf{q}_1) \right) \mathrm{d}\Omega(\mathbf{q}) \\ & \quad = \left(\int_{\mathbb{S}^{m-1}} \left| \omega_{\mathbf{q},\mathbf{q}_2}(A, B) \right| \mathrm{d}\Omega(\mathbf{q}_2) + \int_{\mathbb{S}^{m-1}} \left| -\omega_{\mathbf{q},\mathbf{q}_1}(A, B) \right| \mathrm{d}\Omega(\mathbf{q}_1) \right) \mathrm{d}\Omega(\mathbf{q}) \\ & \quad = \left(2 \int_{\mathbb{S}^{m-1}} \left| \omega_{\mathbf{q},\mathbf{q}_2}(A, B) \right| \mathrm{d}\Omega(\mathbf{q}_2) \right) \mathrm{d}\Omega(\mathbf{q}). \end{aligned}$$

Thus, the derivative of $\mathcal{E}(A, B)$ with respect to $\rho_{\mathbf{w}_1}^A$ is

$$\begin{aligned} \left. \frac{\partial \mathcal{E}(A, B)}{\partial \rho_{\mathbf{w}_1}^A} \right|_{A=B} & = \frac{\partial}{\partial \rho_{\mathbf{w}_1}^A} \left(2 \int_{\mathbb{S}^{m-1}} \left| \omega_{\mathbf{w}_1,\mathbf{q}_2}(A, B) \right| \mathrm{d}\Omega(\mathbf{q}_2) \right) \bigg|_{A=B} \cdot \mathrm{d}\Omega(\mathbf{w}_1) \\ & = 2 \int_{\mathbb{S}^{m-1}} \frac{\partial}{\partial \rho_{\mathbf{w}_1}^A} \left| \arctan \sqrt{\frac{\rho_{\mathbf{q}_2}^A}{\rho_{\mathbf{w}_1}^A}} - \arctan \sqrt{\frac{\rho_{\mathbf{q}_2}^B}{\rho_{\mathbf{w}_1}^B}} \right| \mathrm{d}\Omega(\mathbf{q}_2) \bigg|_{A=B} \cdot \mathrm{d}\Omega(\mathbf{w}_1) \\ & = \int_{\mathbb{S}^{m-1}} \frac{\sqrt{\rho_{\mathbf{q}_2}^A}}{\rho_{\mathbf{w}_1}^A + \rho_{\mathbf{q}_2}^A} \cdot \frac{1}{\sqrt{\rho_{\mathbf{w}_1}^A}} \mathrm{d}\Omega(\mathbf{q}_2) \bigg|_{A=B} \cdot \mathrm{d}\Omega(\mathbf{w}_1) \\ & = \int_{\mathbb{S}^{m-1}} \frac{\sqrt{\rho_{\mathbf{q}_2}^B}}{\rho_{\mathbf{w}_1}^B + \rho_{\mathbf{q}_2}^B} \cdot \frac{1}{\sqrt{\rho_{\mathbf{w}_1}^B}} \mathrm{d}\Omega(\mathbf{q}_2) \cdot \mathrm{d}\Omega(\mathbf{w}_1) \\ & \triangleq \int_{\mathbb{S}^{m-1}} \tau\left(\rho_{\mathbf{w}_1}^B, \rho_{\mathbf{q}_2}^B\right) \mathrm{d}\Omega(\mathbf{q}_2) \cdot \mathrm{d}\Omega(\mathbf{w}_1). \end{aligned}$$

Since

$$\frac{\tau(\rho^B_{\mathbf{w}_1},\rho^B_{\mathbf{q}_2})}{\tau(\rho^B_{\mathbf{w}_2},\rho^B_{\mathbf{q}_2})} = \frac{\frac{\sqrt{\rho^B_{\mathbf{q}_2}}}{\rho^B_{\mathbf{w}_1}+\rho^B_{\mathbf{q}_2}}\cdot\frac{1}{2\sqrt{\rho^B_{\mathbf{w}_1}}}}{\frac{\sqrt{\rho^B_{\mathbf{q}_2}}}{\rho^B_{\mathbf{w}_2}+\rho^B_{\mathbf{q}_2}}\cdot\frac{1}{2\sqrt{\rho^B_{\mathbf{w}_2}}}} = \frac{(\rho^B_{\mathbf{w}_2}+\rho^B_{\mathbf{q}_2})\sqrt{\rho^B_{\mathbf{w}_2}}}{(\rho^B_{\mathbf{w}_1}+\rho^B_{\mathbf{q}_2})\sqrt{\rho^B_{\mathbf{w}_1}}}$$

and $\rho^B_{\mathbf{q}} \geq 0$, $\forall\mathbf{q}$, we have

$$\begin{aligned} \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{0.5} &< \frac{\tau(\rho^B_{\mathbf{w}_1},\rho^B_{\mathbf{q}_2})}{\tau(\rho^B_{\mathbf{w}_2},\rho^B_{\mathbf{q}_2})} \leq \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{1.5}, \quad \text{if } \rho^B_{\mathbf{w}_1} < \rho^B_{\mathbf{w}_2}; \\ \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{1.5} &\leq \frac{\tau(\rho^B_{\mathbf{w}_1},\rho^B_{\mathbf{q}_2})}{\tau(\rho^B_{\mathbf{w}_2},\rho^B_{\mathbf{q}_2})} < \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{0.5}, \quad \text{if } \rho^B_{\mathbf{w}_1} > \rho^B_{\mathbf{w}_2}. \end{aligned}$$

Since the integration is symmetric for all directions, the solid angle elements are all equal and thus $\mathrm{d}\Omega(\mathbf{w}_1) = \mathrm{d}\Omega(\mathbf{w}_2)$. Using Lemma 10, we can obtain that

$$\begin{aligned} \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{0.5} &< \frac{\int_{\mathbb{S}^{m-1}}\tau(\rho^B_{\mathbf{w}_1},\rho^B_{\mathbf{q}_2})\mathrm{d}\Omega(\mathbf{q}_2)}{\int_{\mathbb{S}^{m-1}}\tau(\rho^B_{\mathbf{w}_2},\rho^B_{\mathbf{q}_2})\mathrm{d}\Omega(\mathbf{q}_2)} = \frac{\frac{\partial\mathcal{E}(A,B)}{\partial\rho^A_{\mathbf{w}_1}}|_{A=B}}{\frac{\partial\mathcal{E}(A,B)}{\partial\rho^A_{\mathbf{w}_2}}|_{A=B}} < \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{1.5}, \quad \text{if } \rho^B_{\mathbf{w}_1} < \rho^B_{\mathbf{w}_2}; \\ \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{1.5} &< \frac{\int_{\mathbb{S}^{m-1}}\tau(\rho^B_{\mathbf{w}_1},\rho^B_{\mathbf{q}_2})\mathrm{d}\Omega(\mathbf{q}_2)}{\int_{\mathbb{S}^{m-1}}\tau(\rho^B_{\mathbf{w}_2},\rho^B_{\mathbf{q}_2})\mathrm{d}\Omega(\mathbf{q}_2)} = \frac{\frac{\partial\mathcal{E}(A,B)}{\partial\rho^A_{\mathbf{w}_1}}|_{A=B}}{\frac{\partial\mathcal{E}(A,B)}{\partial\rho^A_{\mathbf{w}_2}}|_{A=B}} < \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{0.5}, \quad \text{if } \rho^B_{\mathbf{w}_1} > \rho^B_{\mathbf{w}_2}. \end{aligned}$$

Then due to the continuity of exponential function, the inequalities can be reformulated as

$$\exists 0.5 < \alpha < 1.5, \quad \text{s.t.}\ \frac{\frac{\partial\mathcal{E}(A,B)}{\partial\rho^A_{\mathbf{w}_1}}|_{A=B}}{\frac{\partial\mathcal{E}(A,B)}{\partial\rho^A_{\mathbf{w}_2}}|_{A=B}} = \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{\alpha},$$

and thus

$$\frac{\mathcal{E}(A_1,B)}{\mathcal{E}(A_2,B)} \approx \frac{\mathcal{E}(B,B) + \frac{\partial\mathcal{E}(A,B)}{\partial\rho^A_{\mathbf{w}_1}}|_{A=B}\cdot\Delta\rho}{\mathcal{E}(B,B) + \frac{\partial\mathcal{E}(A,B)}{\partial\rho^A_{\mathbf{w}_2}}|_{A=B}\cdot\Delta\rho} = \left(\frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}}\right)^{\alpha},$$

where $0.5 < \alpha < 1.5$, and this completes the proof. □

Then, we study the Bregman divergence of scales as the discrepancy and prove Proposition 2.

Proof of Proposition 2 The difference of scales in any direction $\Delta\rho = \rho^A_{\mathbf{w}_i} - \rho^B_{\mathbf{w}_i}$ brings about the discrepancy $\mathcal{D}_\varphi(A,B)$. Since A_1 and A_2 differ from B by $\rho^{A_1}_{\mathbf{w}_1} - \rho^B_{\mathbf{w}_1} = \rho^{A_2}_{\mathbf{w}_2} - \rho^B_{\mathbf{w}_2} = \Delta\rho$, the ratio of dissimilarities has the following relationship (we also have $\mathrm{d}\Omega(\mathbf{w}_1) = \mathrm{d}\Omega(\mathbf{w}_2)$ due to the symmetry of the solid angle elements).

For $\mathcal{D}_{\mathrm{Eu}}(A,B)$, it is straightforward to obtain that

$$\frac{\mathcal{D}_{\mathrm{Eu}}(A_1,B)}{\mathcal{D}_{\mathrm{Eu}}(A_2,B)} = \frac{(\rho^{A_1}_{\mathbf{w}_1} - \rho^B_{\mathbf{w}_1})^2\cdot\mathrm{d}\Omega(\mathbf{w}_1)}{(\rho^{A_2}_{\mathbf{w}_2} - \rho^B_{\mathbf{w}_2})^2\cdot\mathrm{d}\Omega(\mathbf{w}_2)} = \left(\frac{\Delta\rho}{\Delta\rho}\right)^2 = 1.$$

For $D_{\mathrm{KL}}(A, B)$, the result is not straightforwardly available. Since $\Delta\rho$ is assumed to be relatively small, we can estimate it using l'Hospital's rule (Chatterjee 2005) as

$$\begin{aligned}
\frac{\mathcal{D}_{\mathrm{KL}}(A_1, B)}{\mathcal{D}_{\mathrm{KL}}(A_2, B)} &\approx \lim_{\Delta\rho\to 0} \frac{D_{\mathrm{KL}}(\rho^B_{\mathbf{w}_1} + \Delta\rho, \rho^B_{\mathbf{w}_1}) \cdot \mathrm{d}\Omega(\mathbf{w}_1)}{D_{\mathrm{KL}}(\rho^B_{\mathbf{w}_2} + \Delta\rho, \rho^B_{\mathbf{w}_2}) \cdot \mathrm{d}\Omega(\mathbf{w}_2)} \\
&= \lim_{\Delta\rho\to 0} \frac{(D_{\mathrm{KL}}(\rho^B_{\mathbf{w}_1} + \Delta\rho, \rho^B_{\mathbf{w}_1}))'}{(D_{\mathrm{KL}}(\rho^B_{\mathbf{w}_2} + \Delta\rho, \rho^B_{\mathbf{w}_2}))'} = \lim_{\Delta\rho\to 0} \frac{\log(\rho^B_{\mathbf{w}_1} + \Delta\rho) - \log\rho^B_{\mathbf{w}_1}}{\log(\rho^B_{\mathbf{w}_2} + \Delta\rho) - \log\rho^B_{\mathbf{w}_2}} \\
&= \lim_{\Delta\rho\to 0} \frac{(\log(\rho^B_{\mathbf{w}_1} + \Delta\rho) - \log\rho^B_{\mathbf{w}_1})'}{(\log(\rho^B_{\mathbf{w}_2} + \Delta\rho) - \log\rho^B_{\mathbf{w}_2})'} = \lim_{\Delta\rho\to 0} \frac{1/(\rho^B_{\mathbf{w}_1} + \Delta\rho)}{1/(\rho^B_{\mathbf{w}_2} + \Delta\rho)} \\
&= \frac{\rho^B_{\mathbf{w}_2}}{\rho^B_{\mathbf{w}_1}},
\end{aligned}$$

where $f' \triangleq \frac{\mathrm{d}f}{\mathrm{d}\Delta\rho}$ denotes the derivative of f with respect to $\Delta\rho$. □

A.4 The procedure to estimate PG using randomly generated samples

Due to the complexity of its original definition of PG, we resort to the frequency of relative distance consistency as an estimation of PG. In this section, we take the example in Fig. 3 to show the procedure to calculate the estimation of PG using randomly generated samples:

1. Generate 50 sample $\{\mathbf{x}_i\}_{i=1}^{50}$ following a Gaussian distribution, which are shown in each figure.
2. For each metric $M = B, A_1, A_2$, calculate the distance of each pair of points $\{d_M(\mathbf{x}_i, \mathbf{x}_j)\}_{i,j=1,i<j}^{50}$. Then we obtain $C_{50}^2 = 1225$ distances in total and list them as $\{d_M^{(k)}\}_{k=1}^{1225}$. Note that for any M, $d_M^{(k)}$ with the same k corresponds to the same pair of points.
3. To estimate $\mathrm{PG}(d_{A_1}, d_B)$, for any pair of distances (corresponding two pairs of points) $d_M^{(k)}, d_M^{(l)}$, compare their relative distance measured by A_1 and B respectively, i.e., to compare which is greater between $d_{A_1}^{(k)}$ and $d_{A_1}^{(l)}$, then between $d_B^{(k)}$ and $d_B^{(l)}$. Then count the occurrences that A_1 and B give the same judgement, i.e., $d_{A_1}^{(k)} > d_{A_1}^{(l)} \wedge d_B^{(k)} > d_B^{(l)}$ or $d_{A_1}^{(k)} < d_{A_1}^{(l)} \wedge d_B^{(k)} < d_B^{(l)}$ or $d_{A_1}^{(k)} = d_{A_1}^{(l)} \wedge d_B^{(k)} = d_B^{(l)}$. Denoting this count as $N(A_1, B)$, we obtain $N(A_1, B) = 1{,}485{,}043$ and analogously, $N(A_2, B) = 1{,}404{,}431$.
4. Since there are totally $1{,}225^2$ possible pairs of distances, the frequencies of the event "geometry preserved" are $\mathrm{PG}_f(A_1, B) \approx N(A_1, B)/1{,}225^2 = 0.990$ and $\mathrm{PG}_f(A_1, B) \approx N(A_2, B)/1{,}225^2 = 0.936$.

It is notable that although the samples are generated following a Gaussian distribution, it does not imply the f in $\mathrm{PG}_f(A_1, B)$ is Gaussian. Indeed, in the procedure above, the two pairs of the points are not independent and each pair is sampled following a uniform distribution, which makes f more complex.

References

Argyriou, A., & Evgeniou, T. (2008). Convex multi-task feature learning. *Machine Learning*, *73*(3), 243–272.

Argyriou, A., Micchelli, C. A., Pontil, M., & Ying, Y. (2008). A spectral regularization framework for multi-task structure learning. *Advances in Neural Information Processing Systems*, *20*, 25–32.

Banerjee, A., Merugu, S., Dhillon, I. S., & Ghosh, J. (2005). Clustering with Bregman divergences. *Journal of Machine Learning Research*, *6*, 1705–1749.

Bauschke, H. H., & Borwein, J. M. (2001). Joint and separate convexity of the Bregman distance. In *Inherently parallel algorithms in feasibility and optimization and their applications* (Vol. 8, p. 23–36). Amsterdam: North-Holland.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. New York: Cambridge University Press.

Burago, D., Burago, Y., & Ivanov, S. (2001). *A course in metric geometry*. Providence: Am. Math. Soc.

Caruana, R. (1997). Multitask learning. *Machine Learning*, *28*(1), 41–75.

Chatterjee, D. (2005). *Real analysis*. New York: Prentice Hall.

Chung, F. R. K. (1997). *CBMS regional conference series in mathematics: Vol. 92*. *Spectral graph theory*. Providence: Am. Math. Soc.

Davis, J. V., Kulis, B., Jain, P., Sra, S., & Dhillon, I. S. (2007). Information-theoretic metric learning. In *Proceedings of the 24th international conference on machine learning* (pp. 209–216).

Dhillon, I. S., & Tropp, J. A. (2008). Matrix nearness problems with Bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, *29*, 1120–1146.

Evgeniou, T., & Pontil, M. (2004). Regularized multi-task learning. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 109–117).

Friedman, J., Hastie, T., Hofling, H., & Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, *1*(2), 302–332.

Gupta, A. K., & Nagar, D. K. (2000). *Matrix variate distributions* (Vol. 104). London: Chapman & Hall.

Hardy, G. H., Littlewood, J. E., & Polya, G. (1988). *Inequalities* (2nd ed.). Cambridge: Cambridge University Press.

Horn, R., & Johnson, C. (1985). *Matrix analysis*. Cambridge: Cambridge University Press.

Horn, R. A., & Johnson, C. R. (1991). *Topics in matrix analysis*. Cambridge: Cambridge University Press.

Huang, K., Ying, Y., & Campbell, C. (2009). Gsml: a unified framework for sparse metric learning. In *Ninth IEEE international conference on data mining* (pp. 189–198).

Huang, K., Ying, Y., & Campbell, C. (2011). Generalized sparse metric learning with relative comparisons. *Knowledge and Information Systems*, *28*(1), 25–45.

Jin, R., Wang, S., & Zhou, Y. (2009). Regularized distance metric learning: theory and algorithm. In *Advances in neural information processing systems* (Vol. 22, pp. 862–870).

Kingravi, H. A. (2007). On high dimensional spaces, and an issue they pose in machine learning. http://www.cc.gatech.edu/~kingravi/notes.html.

Kulis, B., Sustik, M. A., & Dhillon, I. S. (2009). Low-rank kernel learning with Bregman matrix divergences. *Journal of Machine Learning Research*, *10*, 341–376.

Lewis, A. S. (1996). Convex analysis on the Hermitian matrices. *SIAM Journal on Optimization*, *6*, 164–177.

Lindblad, G. (1973). Entropy, information and quantum measurements. *Communications in Mathematical Physics*, *33*(4), 305–322.

Lindblad, G. (1975). Completely positive maps and entropy inequalities. *Communications in Mathematical Physics*, *40*(2), 147–151.

Marshall, A., Olkin, I., & Arnold, B. (2011). *Inequalities: theory of majorization and its applications*. Berlin: Springer.

Nielsen, M. A., & Chuang, I.L. (2010). *Quantum computation and quantum information*. New York: Cambridge University Press.

Parameswaran, S., & Weinberger, K. (2010). Large margin multi-task metric learning. *Advances in Neural Information Processing Systems*, *23*, 1867–1875.

Tropp, J. A. (2012). From joint convexity of quantum relative entropy to a concavity theorem of Lieb. *Proceedings of the American Mathematical Society*, *140*, 1757–1760.

Tseng, P. (1988). *Coordinate ascent for maximizing nondifferentiable concave functions* (Tech. Rep.). Center for Intelligent Control Systems (US) and Massachusetts Institute of Technology, Laboratory for Information and Decision Systems.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, *109*, 475–494.

Watrous, J. (2008). Lecture notes: theory of quantum information. http://www.cs.uwaterloo.ca/~watrous/quant-info/.

Weinberger, K. Q., & Saul, L. K. (2008). Fast solvers and efficient implementations for distance metric learning. In *Proceedings of the 25th international conference on machine learning* (pp. 1160–1167).

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, *10*, 207–244.

Wikipedia (2012). Characteristic function (convex analysis)—wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Characteristic_function_(convex_analysis)&oldid=490880941.

Wikipedia (2013a). N-sphere—wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=N-sphere&oldid=549291786.

Wikipedia (2013b). Semi-continuity—wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Semi-continuity&oldid=540746717.

Xing, E. P., Ng, A. Y., Jordan, M. I., & Russell, S. (2003). Distance metric learning, with application to clustering with side-information. In *Advances in neural information processing systems* (pp. 505–512).

Yang, P., Huang, K., & Liu, C.-L. (2011). Multi-task low-rank metric learning based on common subspace. In *Proceedings of the 18th international conference on neural information processing—volume part ii* (Vol. 7063, pp. 151–159).

Yang, P., Huang, K., & Liu, C.-L. (2012). Geometry preserving multi-task metric learning. In *European conference on machine learning and knowledge discovery in databases* (Vol. 7523, pp. 648–664).

Ying, Y., & Li, P. (2012). Distance metric learning with eigenvalue optimization. *Journal of Machine Learning Research*, *13*, 1–26.

Ying, Y., Huang, K., & Campbell, C. (2009). Sparse metric learning via smooth optimization. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems 22* (pp. 2214–2222).

Zhang, Y., & Yeung, D.-Y. (2010a). A convex formulation for learning task relationships in multi-task learning. In *Proceedings of the proceedings of the twenty-sixth conference annual conference on uncertainty in artificial intelligence* (pp. 733–742).

Zhang, Y., & Yeung, D.-Y. (2010b). Transfer metric learning by learning task relationships. In *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining*.

Zhang, J., Ghahramani, Z., & Yang, Y. (2008). Flexible latent variable models for multi-task learning. *Machine Learning*, *73*(3), 221–242.