Self-Supervised Metric Learning in Multi-View Data: A Downstream Task Perspective

Shulei Wang

Abstract

Self-supervised metric learning has been a successful approach for learning a distance from an unlabeled dataset. The resulting distance is broadly useful for improving various distance-based downstream tasks, even when no information from downstream tasks is utilized in the metric learning stage. To gain insights into this approach, we develop a statistical framework to theoretically study how self-supervised metric learning can benefit downstream tasks in the context of multi-view data. Under this framework, we show that the target distance of metric learning satisfies several desired properties for the downstream tasks. On the other hand, our investigation suggests the target distance can be further improved by moderating each direction's weights. In addition, our analysis precisely characterizes the improvement by self-supervised metric learning on four commonly used downstream tasks: sample identification, two-sample testing, k-means clustering, and k-nearest neighbor classification. When the distance is estimated from an unlabeled dataset, we establish the upper bound on distance estimation's accuracy and the number of samples sufficient for downstream task improvement. Finally, numerical experiments are presented to support the theoretical results in the paper.

Keywords: Metric learning; k-means; k-nearest neighbor; Two-sample testing

Shulei Wang is an Assistant Professor, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820 (E-mail: *shuleiw@illinois.edu*).

1 Introduction

1.1 Self-Supervised Metric Learning in Multi-View Data

Measuring distance is the first step to understand relationships between the data points and also one of the most key components in many distance-based statistics and machine learning methods, such as the k-means clustering algorithm and k-nearest neighbor method. The performance of these distance-based methods usually depends in large part on the choice of distance. Although various distances have been proposed to quantify the difference between data points in different applications, e.g., Euclidean distance, Wasserstein distance, and Manhattan distance, it is still unclear which distance the researcher should use to quantify the dissimilarity between the data for a given task at hand. One promising solution for such a problem is metric learning, which has already been used in a wide range of applications, including face identification (Guillaumin, Verbeek, and Schmid, 2009; Liao et al., 2015; Li et al., 2014; Yi et al., 2014), remote sensing (Zhang, Lu, and Li, 2018; Ji et al., 2018) and neuroscience (Ktena et al., 2018; Ma et al., 2019).

Most metric learning methods require access to similar and dissimilar data pairs since they aim to preserve the closeness between similar data pairs and push dissimilar data points far from each other. A commonly-used strategy is to construct similar and dissimilar data pairs based on the labels' value in a supervised setting. For example, when the label is binary, the data points within the same class are regarded as similar ones, and those from different classes are dissimilar ones. Despite of the popularity in practice, such a strategy usually needs a large amount of labeled data, which can sometimes be expensive or difficult to collect. To overcome this challenge, a self-supervised learning framework is proposed to leverage the unlabeled data (Zhang, Isola, and Efros, 2016; Oord, Li, and Vinyals, 2018; Tian, Krishnan, and Isola, 2019; Chen et al., 2020a). The pseudo labels are generated from the unlabeled dataset itself, and then the statistics or machine learning model is trained by these pseudo labels. Specifically, when it comes to self-supervised metric learning, similar and dissimilar data pairs are constructed in an unsupervised fashion from the unlabeled dataset to train a better distance.

It is generally difficult to distinguish similar and dissimilar data pairs from unstructured data as we usually do not have insights on which data points are closer than which. However, it can be much easier to construct similar pairs in an unsupervised way when there is some structure information in the dataset. In particular, multi-view data is a typical class of such datasets, where several different views from each sample are observed. More concretely, multi-view data refers to a dataset of m samples, in which n different views of each sample $(X_{i,1}, \ldots, X_{i,n}) \in \mathbb{R}^{d \times n}$, $i = 1, \ldots, m$, are recorded. Multi-view data is very common in real applications, for instance:

- In face recognition, the images of the same face with different illumination or viewpoints are collected, such as the Extended Yale Face Database B (Georghiades, Belhumeur, and Kriegman, 2001).

- In the microbiome studies, the microbial samples of the same individual are usually collected at multiple time points (Gajer et al., 2012; Flores et al., 2014).

- In robotics, the videos of the same scenario from multiple viewpoints are recorded (Sermanet et al., 2017; Dwibedi et al., 2018).

- Data augmentation is a popular technique to help increase the amount of data and generate extra views for each sample. For example, many different ways are used to synthesize imaging data, such as flipping, rotation, colorization, and cropping (Gidaris, Singh, and Komodakis, 2018; Shorten and Khoshgoftaar, 2019). By the data augmentation technique, a multi-view dataset can be generated from a single-view dataset.

In these multi-view datasets, one can naturally label data points from two different views of the same sample, $X_{i,j}$ and $X_{i,j'}$ for some $j \neq j'$, as similar pair and data points from different samples, $X_{i,j}$ and $X_{i',j'}$ for some $i \neq i'$, as dissimilar pair. Therefore, it is a popular strategy to use multi-view data for self-supervised metric learning, which has been very successful in practice (Sohn, 2016; Movshovitz-Attias et al., 2017; Sermanet et al., 2017; Duan et al., 2018; Tian, Krishnan, and Isola, 2019; Roth et al., 2020; Deng et al., 2021).

Given the similar and dissimilar data pairs, a common principle of most existing metric learning methods is to look for a distance that can better predict whether a pair of data points is similar or not. If similar and dissimilar data pairs come from the multi-view data, it is equivalent to find a distance that can distinguish if a pair of data points comes from the same sample or not. To achieve this goal, different loss functions have been proposed to compare data pairs in metric learning (Xing et al., 2002; Weinberger and Saul, 2009; Kulis, 2012; Bellet, Habrard, and Sebban, 2013, 2015; Musgrave, Belongie, and Lim, 2020). Despite the difference in these loss functions, the ideal distance in metric learning methods aims to have a much larger value for dissimilar data pairs than similar ones.

1.2 Self-Supervised Metric Learning and Downstream Task

Learning a distance from multi-view data is never the end of story, and the ultimate goal of self-supervised metric learning is to improve various downstream distance-based methods, be it k-means clustering algorithm or k-nearest neighbor method. In the supervised setting, where similarity is determined based on the actual labels, it is natural to believe that the resulting distance from metric learning can benefit the downstream tasks since similar and dissimilar data pairs are directly related to the labels in the downstream analysis (Weinberger and Saul, 2009). On the other hand, different from the supervised setting, the self-supervised metric learning only has access to the fact whether two data points come from the same sample or not. At first sight, the self-supervised metric learning seems impossible to improve the performance of downstream distance-based methods since it does not utilize any label information. However there is considerable empirical evidence showing that self-supervised metric learning can indeed improve the efficiency of downstream analysis (Schroff, Kalenichenko, and Philbin, 2015; Sermanet et al., 2017; Tian, Krishnan, and Isola, 2019). These phenomena raise several natural questions: why does self-supervised metric learning benefit the downstream tasks? What kind of distance is a reasonable distance from an angle of downstream analysis? To what extent can the downstream tasks be improved by self-

supervised metric learning? How much unlabeled multi-view data is sufficient to help improve the downstream tasks?

The theoretical properties of metric learning are mainly studied from the angle of generalization rates under a supervised setting in the literature (Jin, Wang, and Zhou, 2009; Bellet, Habrard, and Sebban, 2015; Cao, Guo, and Ying, 2016; Jain, Mason, and Nowak, 2017; Ye, Zhan, and Jiang, 2019). These results could help us understand how fast the empirical loss function converges but do not connect the resulting distance with downstream tasks. On the other hand, the self-supervised metric learning we study here is closely connected with self-supervised representation learning, which aims to find a transformation of the data that makes it easier to build an efficient classifier (Bengio, Courville, and Vincent, 2013; Tschannen et al., 2019). Instead of distance, some recent works study how the representation learned from the data is helpful for the downstream tasks under a self-supervised setting (Arora et al., 2019; Lee et al., 2020; Tian et al., 2020; Tosh, Krishnamurthy, and Hsu, 2021; Wei et al., 2020; Tsai et al., 2020). Although these results provide theoretical insights of self-supervised representation learning, the analysis cannot be directly applied to the investigation of metric learning and the downstream distance-based task, such as k-means clustering algorithm and k-nearest neighbor method. Therefore, there is a clear need for a comprehensive theoretical study for self-supervised metric learning from a perspective of the downstream task.

1.3 A Downstream Task Perspective

This paper's main goal is to understand how self-supervised metric learning works from the perspective of the downstream task. To demystify the effectiveness of self-supervised metric learning, we focus on learning a Mahalanobis distance, which has the form $D_M(X_1, X_2) = (X_1 - X_2)^T M(X_1 - X_2)$ for some positive semi-definite matrix M, and assume the multi-view data $(X_{i,1}, \ldots, X_{i,n})$ is drawn from a latent factor model

$$X_{i,j} = BZ_i + \epsilon_{i,j}, \qquad j = 1, \ldots, n,\ i = 1, \ldots, m$$

where $Z_i \in \mathbb{R}^K$ is ith sample's unobserved latent variable and $B = (b_1, \ldots, b_K)$ is the collection of factors such that $B^T B = \Lambda$, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ is a diagonal matrix. Here, $\epsilon_{i,j}$ is some view-specific random variable independent from Z_i. Under this latent factor model, the intrinsic structure of data lies in a K-dimensional subspace, where K is usually much smaller than d. Our investigation shows that the target distances of metric learning under the latent factor model can be seen as the following distance

$$D^*(X_1, X_2) = (X_1 - X_2)^T B B^T (X_1 - X_2).$$

Roughly speaking, the target distance D^* measures the difference between data within the K-dimensional subspace spanned by $b_1, \ldots, b_K$ and puts more weights in the directions that can better distinguish the similar and dissimilar data pairs. Thus, the distance can help reduce the data dimension, but is this distance a reasonable distance for downstream analysis?

The target distance D^* seems only related to the latent factor model of multi-view data and has nothing to do with downstream tasks. However, our analysis shows that, perhaps surprisingly, D^* has several desired properties for the downstream tasks if we further assume the latent variable includes all the label information in the downstream analysis, i.e.,

$$Y_i \perp (X_{i,1}, \ldots, X_{i,n}) | Z_i,$$

where $Y_i \in \{-1, 1\}$ is the binary label in the downstream analysis. Here, no assumption is made for the relationship between label Y and latent variable Z. Specifically, the distance D^* has the following properties: 1) D^* is a distance between a sufficient statistic for Y, so no information on the label is lost; 2) D^* is robust to a collection of spurious features in data; 3) D^* only keeps minimally sufficient information for Y. In a word, the distance that self-supervised metric learning aims for can help remove nuisance factors and keep necessary information even when no label is utilized. On the other hand, our further analysis suggests that the directions that can better capture the difference between the similar and dissimilar data pairs are not necessarily more useful in the downstream tasks than the one that cannot capture the difference very well. Motivated by this observation, we argue that target distance D^* can be improved by an isotropic version of target

distance, that is, we put equal weights in all directions

$$D^{**}(X_1, X_2) = (X_1 - X_2)^T B \Lambda^{-1} B^T (X_1 - X_2).$$

In particular, our results indicate that the distance D^{**} is a better choice than D^* when the condition number of factor model is large where condition number is defined as $\kappa = \lambda_1/\lambda_K$.

Downstream Task	Measure	Euclidean Distance	Metric Learning
sample identification	detection radius	$\frac{d^{1/4}\sigma}{\sqrt{\lambda}}$	$\frac{K^{1/4}\sigma}{\sqrt{\lambda}}$
two-sample test	detection radius	$\left(\frac{\sqrt{K}\lambda + \sqrt{d}\sigma^2}{s}\right)^{1/2}$	$\left(\frac{\sqrt{K}(\lambda + \sigma^2)}{s}\right)^{1/2}$
k-means	mis-cluster rate	$\exp\left(-\frac{\|\mu\|^2}{8(\lambda+\sigma^2)}\right)$	$\exp\left(-\frac{\|\mu\|^2}{8(\lambda+\sigma^2)}\right)$
	required signal	$\left(1+\frac{K}{s}\right)\lambda + \left(1+\frac{d}{s}\right)\sigma^2$	$\left(1+\frac{K}{s}\right)(\lambda+\sigma^2)$
k-nearest neighbor	excess risk	$s^{-\alpha(1+\beta)/(2\alpha+d)}$	$s^{-\alpha(1+\beta)/(2\alpha+K)}$

Table 1: Performance comparisons between Euclidean distance and resulting distance from self-supervised metric learning. d is the dimension of the data, K is the number of factors, s is the sample size in the downstream task, σ^2 measures the variation of different views, λ measures the variation of sample difference, and μ is the expected difference between class.

To further investigate the benefits of self-supervised metric learning, we compare the performance of Euclidean distance and target distances from metric learning, both D^* and D^{**}, on four commonly used distance-based methods: distance-based sample identification, distance-based two-sample testing, k-means clustering, and k-nearest neighbor (k-NN) classification algorithm. The informal results are summarized in Table 1 if we assume $\lambda = \lambda_1 = \ldots = \lambda_K$ and the covariance matrix of $\epsilon_{i,j}$ is $\sigma^2 I$. The formal results of a general setup, including both upper and lower bound, are discussed in Section 4. Table 1 suggests that the performance of downstream tasks can be improved in different ways. In particular, the curse of dimensionality can be much alleviated by self-supervised metric learning as the performance only relies on the number of factors K rather than the dimension of data d when self-supervised metric learning is applied. For example, the

nonparametric method k-NN behaves just like on a K-dimensional space as the target distance D^* and D^{**} fits the geometry of the Bayes classification rule in a better way.

Downstream Task	Distance	Accuracy Δ	Sample Size m
two-sample test k-means sample identification	D^*	$o(\lambda)$	$K + \frac{d\sigma^2}{n\lambda} + \frac{d\sigma^4}{n^2\lambda^2}$
	D^{**}	$o(1)$	$\frac{d\sigma^2}{n\lambda} + \frac{d\sigma^4}{n^2\lambda^2}$
k-nearest neighbor	D^*	$\lambda s^{-1/(2\alpha+K)}$	$s^{1/(2\alpha+K)}\left(K + \frac{d\sigma^2}{n\lambda} + \frac{d\sigma^4}{n^2\lambda^2}\right)$
	D^{**}	$s^{-1/(2\alpha+K)}$	$s^{1/(2\alpha+K)}\left(\frac{d\sigma^2}{n\lambda} + \frac{d\sigma^4}{n^2\lambda^2}\right)$

Table 2: Distance estimation's accuracy and number of samples sufficient for downstream task improvement in self-supervised metric learning.

In practice, we still need to estimate the target distances D^* and D^{**} from the unlabeled multi-view data when they are unknown in advance. Our investigation shows that the estimated distances from self-supervised metric learning can also help improve above four distance-based methods provided the distance estimation is accurate enough. Specifically, if we quantify the distance estimation's accuracy by their largest discrepancy

$$\Delta(D, \hat{D}) = \sup_{\|X_1 - X_2\| \leq 1} \left| D(X_1, X_2) - \hat{D}(X_1, X_2) \right|,$$

the sufficient accuracy to achieve results in Table 1 is summarized in Table 2. To estimate an accurate distance for downstream tasks, we consider a spectral metric learning method and study its theoretical properties in this paper. We show that the spectral method can help achieve minimax optimality in estimating target distances. Moreover, the analysis can help precisely characterize the number of samples m sufficient for downstream tasks improvement, which is also summarized in Table 2. Table 2 shows that it is easier to estimate D^{**} than D^* from the unlabeled multi-view data.

The rest of the paper is organized as follows. We first introduce the multi-view model and discuss the main assumptions of the model in Section 2. Next, Section 3 studies the target distance of metric learning methods and its properties from a perspective of downstream analysis. In Sec-

tion 4, the benefits of self-supervised learning are systematically investigated on several specific downstream distance-based tasks. Then, we study target distance estimation and characterize the sample complexity for downstream tasks improvement in Section 5. Finally, we analyze both the simulated and real data sets in Section 6 to verify the theoretical results in this paper. All proofs are relegated to online Supplemental Materials.

2 A Model for Multi-View Data

In this paper, we consider the following model of multi-view data for m different samples

$$(X_{i,1}, \ldots, X_{i,n}, Z_i, Y_i), \qquad i = 1, \ldots m,$$

where n is the number of views observed for each sample. We assume each (Z_i, Y_i) is independently drawn from a distribution $\pi(Z, Y)$, where $Z \in \mathbb{R}^K$ represents the sample's latent variable, and Y is the label of interest. For simplicity, we always assume the label of interest is binary, i.e., $Y \in \{-1, 1\}$. We also assume the conditional distribution of Z given Y is a continuous distribution, that is, the probability density function $\pi(Z|Y)$ exists. Given the latent variable Z_i, we assume the data of n different views $X_{i,j} \in \mathbb{R}^d$, $j = 1, \ldots, n$, are independently drawn from a continuous conditional distribution $f(X|Z)$. In self-supervised metric learning, instead of observing the full data, we only observe the unlabeled multi-view data, i.e.,

$$(X_{i,1}, \ldots, X_{i,n}), \qquad i = 1, \ldots m.$$

In the downstream analysis, depending on the task, we assume the observed data is a collection of single-view data with or without labels, i.e.,

$$(X_1, Y_1), \ldots, (X_s, Y_s) \qquad \text{or} \qquad X_1, \ldots, X_s.$$

Here, X_i refers to the single-view data in downstream analysis, and $X_{i,j}$ refers to the multi-view data in metric learning. We assume the data used in metric learning and downstream analysis are drawn from the same distribution, but different parts of the data are observed. In a typical self-

supervised learning setting, we can expect the sample size in unlabeled multi-view data m is much larger than the sample size in the downstream analysis s.

The latent variable Z plays a vital role in the structure of multi-view data, characterizing the information shared by different views of the same sample. We assume $X_{i,j}$ connects with Z_i through a factor model (Fan et al., 2020), i.e.,

$$X_{i,j} = \sum_{k=1}^{K} b_k Z_{i,k} + \epsilon_{i,j} \tag{1}$$

where $\epsilon_{i,j}$ is a mean zero random variable independent from Z_i. $\epsilon_{i,j}$ are independent for different i and j. If we write $B = (b_1, \ldots, b_K)$, we further assume

$$B^T B = \Lambda \qquad \text{and} \qquad \mathrm{Var}(Z) = I_K,$$

where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_K)$ is a diagonal matrix with $\lambda_1 \geq \ldots \geq \lambda_K$ and I_K is an identity matrix. In addition, we assume $(I_d - B(B^T B)^{-1} B^T)\epsilon_{i,j}$ is independent from $B^T \epsilon_{i,j}$. This latent factor model assumes that the intrinsic structure of data lies in a K-dimensional subspace. In the rest of the paper, we write $U = B\Lambda^{-1/2}$ as normalized projection matrix and $u_k = b_k/\sqrt{\lambda_k}$. Besides, we also assume the latent variable Z includes all information about the sample which is invariant from different views, and thus

$$Y_i \perp (X_{i,1}, \ldots, X_{i,n}) | Z_i. \tag{2}$$

In other words, the observed multi-view data is connected with the label of interest only through the latent variable.

3 Self-Supervised Metric Learning

3.1 Metric Learning

Given the multi-view data, metric learning aims to learn a distance D that can help improve the downstream tasks. In particular, many different loss functions have been proposed to separate similar and dissimilar data pairs in the literature of metric learning Kulis (2012); Musgrave, Belongie, and Lim (2020), including contrastive loss (Xing et al., 2002; Chopra, Hadsell, and LeCun,

2005; Hadsell, Chopra, and LeCun, 2006), the triplet loss (Weinberger and Saul, 2009; Chechik et al., 2010; Schroff, Kalenichenko, and Philbin, 2015), and N-pair loss(Sohn, 2016). These loss functions have been widely used in various applications and lead to good performance in practice.

We now study how metric learning can extract information from the similar and dissimilar data pairs. The common goal of different metric learning methods is to find a distance that can distinguish dissimilar and similar data pairs. This goal can be naturally achieved by maximizing the following expected distance difference between dissimilar and similar data pairs in multi-view data

$$\mathbb{E}\left(D(X_{i,j}, X_{i',j'}) - D(X_{i,j}, X_{i,j'})\right),$$

where $X_{i,j}$ and $X_{i',j'}$ are from different samples, and $X_{i,j}$ and $X_{i,j'}$ are different views of the same sample. If we are interested in learning a Mahalanobis distance, we can show that

$$M^* := \underset{M \in \mathbb{S}_+^{d\times d}, \|M\|_F \le 1}{\operatorname{argmax}} \mathbb{E}\left(D_M(X_{i,j}, X_{i',j'}) - D_M(X_{i,j}, X_{i,j'})\right) = BB^T / \|BB^T\|_F, \tag{3}$$

where $\mathbb{S}_+^{d\times d}$ is the collection of symmetric and positive semi-definite matrix and the Frobenius norm of a matrix M is defined as $\|M\|_F = \sqrt{\sum_{i=1}^d \sigma_i^2(M)}$ where $\sigma_i(M)$ are the singular values of M. The main purpose of constraint for the Frobenius norm of M is to avoid the scaling issue of Mahalanobis distance. For example, we always have $\mathbb{E}\left(D_{cM}(X_{i,j}, X_{i',j'}) - D_{cM}(X_{i,j}, X_{i,j'})\right) > \mathbb{E}\left(D_M(X_{i,j}, X_{i',j'}) - D_M(X_{i,j}, X_{i,j'})\right)$ for any constant $c > 1$. When we observe infinite samples, the target Mahalanobis distance in above metric learning formulation is

$$D^*(X_1, X_2) = (X_1 - X_2)^T BB^T (X_1 - X_2) = (X_1 - X_2)^T U \Lambda U^T (X_1 - X_2).$$

Compared with the Euclidean distance, the target distance D^* makes two main modifications: (i) D^* measures the difference between data points in K directions spanned by the column space of B; (ii) D^* puts different weights in different directions. Is this distance D^* a reasonable distance for the downstream analysis?

3.2 Distance for Downstream Task

The self-supervised metric learning aim to learn a distance D^* by the unlabeled multi-view data. However, it is still unclear how the target distance D^* is linked with the downstream tasks. In this section, we will see that the distance D^* has several good properties desired for the downstream tasks, but may not honestly reflect the information needed for the downstream analysis. To see this, we need the following theorem.

Theorem 1. *Suppose all the assumptions for multi-view data model in Section 2 hold. Then there exists a function g and a vector $\theta \in \mathbb{R}^K$ with $\|\theta\| < 2$ such that*

$$\frac{\pi(X|Y=1)}{\pi(X|Y=-1)} = g(U^T X) \qquad \text{and} \qquad \mathbb{E}(X|Y=1) - \mathbb{E}(X|Y=-1) = B\theta,$$

where U is the normalized projection matrix in factor model and $\pi(X|Y)$ is the probability density function of X given Y. Moreover, for any given $\theta \in \mathbb{R}^K$ with $\|\theta\| < 2$, there exists a joint distribution of (X, Z, Y) satisfying assumptions in Section 2 such that

$$\mathbb{E}(X|Y=1) - \mathbb{E}(X|Y=-1) = B\theta.$$

Theorem 1 shows that D^* has the following good properties for downstream tasks:

- In Theorem 1, it is shown that $U^T X$ is a sufficient statistic for Y. Thus, from a prediction view, no information on Y is lost when D^* is used. This property is also a gold standard of many other problems, including approximate Bayesian computation (Fearnhead and Prangle, 2012), representation learning (Cvitkovic and Koliander, 2019), and dimension reduction (Adragni and Cook, 2009).

- Theorem 1 suggests the mean difference between classes lies in the column space of B. If we write $U_\perp$ as an orthogonal matrix of U, then $U_\perp^T X$ is a collection of spurious features. D^* is robust to these spurious features.

- As suggested by the second part of Theorem 1, all $u_k^T X$, $k = 1, \ldots, K$, are potentially useful when we do not have access to Y in the metric learning stage. In other words, the distance D^* only keeps minimally sufficient information of X for Y.

In a word, the distance D^* can keep all necessary information for Y and remove nuisance factors from the data X, although label information is not utilized in the metric learning stage.

Unlike Euclidean distance, the target distance D^* puts more weights in the directions that can reflect more difference between similar and dissimilar data pairs. More concretely, if we project the data to the direction u_k, the difference between similar and dissimilar data pairs is λ_k

$$\lambda_k = \mathbb{E}\left(\left[u_k^T(X_{i,j} - X_{i',j'})\right]^2 - \left[u_k^T(X_{i,j} - X_{i,j'})\right]^2\right), \qquad k = 1, \ldots, K.$$

Along direction u_k, the average distance between dissimilar data pair is more significant than that between similar data pair when λ_k is larger. So u_k can better distinguish similar and dissimilar data pairs than u_{k+1} as $\lambda_k \geq \lambda_{k+1}$. It seems reasonable to put more weights on u_k over u_{k+1} since it is usually believed that a feature that can better distinguish similar and dissimilar data pairs is more useful for the downstream analysis. However, the second part of Theorem 1 suggests that it is possible that u_{k+1} is more useful than u_k in the downstream analysis. For example, if we assume $Z|Y \sim N(\theta Y/2, I_K - \theta\theta^T/4)$ with θ such that $\theta_k = 0$ but $\theta_{k+1} \neq 0$, then $u_k^T X|Y = 1$ and $u_k^T X|Y = -1$ follow the same distribution while $u_{k+1}^T X|Y = 1$ and $u_{k+1}^T X|Y = -1$ follow different ones. Motivated by this observation, we consider a moderated target distance

$$D^{**}(X_1, X_2) = (X_1 - X_2)^T U U^T (X_1 - X_2),$$

which puts equal weights in all directions u_k, $k = 1, \ldots, K$. Similar to D^*, D^{**} also has the same good properties for the downstream tasks. As we can see in the next section, D^{**} is a better choice than D^* when the conditional number $\kappa = \lambda_1/\lambda_K$ is large.

4 Target Distance on Specific Tasks

The ultimate goal of self-supervised metric learning is to improve various downstream distance-based statistical and machine learning methods. But it is still unclear to what extent the perfor-

mance of the specific downstream task can be improved. In order to fill this gap, we investigate the benefits of self-supervised metric learning on some specific tasks when we observe infinite unlabeled multi-view samples, that is, D^* and D^{**} are known. We consider four of the most commonly used distance-based methods: k-nearest neighbor classification algorithm, distance-based two-sample testing, k-means clustering (discussed in Supplemental Materials), and distance-based sample identification (discussed in Supplemental Materials).

4.1 k-Nearest Neighbor Classification

Classification is the first problem we consider in this section. The observed data in classification includes the label of each sample, i.e., $(X_1, Y_1), \ldots, (X_s, Y_s)$. In classification, our goal is to build a decision rule $f : \mathbb{R}^d \to \{-1, 1\}$ to predict the label Y for any given input of X. A long list of classification methods has been proposed to predict the labels. One of the most simple, intuitive, and efficient ones is probably the k-nearest neighbor (k-NN) classification method (Fix, 1985; Altman, 1992; Biau and Devroye, 2015). Given the choice of distance D and a fixed point x, k-NN is defined as following: $(X_{(1)}, Y_{(1)}), \ldots, (X_{(s)}, Y_{(s)})$ is a permutation of $(X_1, Y_1), \ldots, (X_s, Y_s)$ such that

$$D(X_{(1)}, x) \leq \ldots \leq D(X_{(s)}, x),$$

and then the decision rule of k-NN is the majority vote of its neighbors

$$\hat{f}_D(x) = \begin{cases} 1, & \sum_{i=1}^{k} \mathbf{I}(Y_{(i)} = 1) \geq k/2, \\ -1, & \text{otherwise.} \end{cases}$$

The k-NN classification rule is a plug-in estimator of the Bayes classification rule, which is given by

$$f^*(x) = \begin{cases} 1, & \eta(x) \geq 1/2, \\ -1, & \text{otherwise,} \end{cases}$$

where $\eta(x) = \mathbb{P}(Y = 1 | X = x)$ is the regression function. The Bayesian rule is considered as the optimal decision rule since it minimizes misclassification error $R(f) = \mathbb{P}(Y \neq f(X))$. To com-

pare the performances of different distances on k-NN, we use the excess risk of misclassification error as the measure

$$r(D) = \mathbb{E}\left(\mathbb{P}(Y \neq \hat{f}_D(X))\right) - \mathbb{P}(Y \neq f^*(X)).$$

Before characterizing the performance of k-NN, we can show that both the Bayes classification rule and the regression function can be written as a function of $U^T X$. A toy example of regression function is shown in Figure 1 to illustrate the idea. The form of the regression function is closely connected to the multiple index model in statistical literature (Li, 1991; Lin et al., 2021).

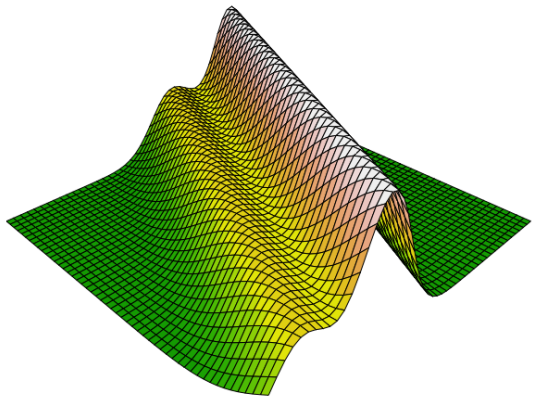

Figure 1: A toy example of regression function in two dimensional space. The regression function only changes along one direction.

Proposition 1. *If the assumptions in Section 2 hold, there exists a function $\tilde{\eta}$ and $\tilde{f}^*$ such that*

$$\eta(x) = \tilde{\eta}(U^T x) \qquad \text{and} \qquad f^*(x) = \tilde{f}^*(U^T x).$$

We omit the proof of Proposition 1 since it is an immediate result of Theorem 1. Proposition 1 suggests that we can make assumptions for $\tilde{\eta}$ and $\tilde{f}^*$ rather than η and f^*. Specifically, we consider the following assumptions.

Assumption 1. *It holds that*

(a) $\tilde{\eta}(y)$ is α-Hölder continuous, i.e., $|\tilde{\eta}(y) - \tilde{\eta}(y')| \leq L\|y - y'\|^\alpha$, where $y, y' \in \mathbb{R}^K$;

(b) the distribution of X satisfies β-marginal assumption, i.e., $\mathbb{P}(0 < |\tilde{\eta}(U^T X) - 1/2| \leq t) \leq C_0 t^\beta$ for some constant C_0;

(c) the support of X is a compact set and the probability density function $\mu(x)$ exists. The probability density function $\mu(x)$ is bounded away from 0 on the support of X, i.e., $\mu(x) \geq \mu_{\min}$ for some small constant $\mu_{\min}$.

These assumptions in Assumption 1 are commonly used conditions for analyzing nonparametric classification methods such as k-NN (Audibert and Tsybakov, 2007; Samworth, 2012). With these conditions, the following theorem characterizes the convergence rate of k-NN when different distances are used.

Theorem 2. *Suppose assumptions in Section 2 and Assumption 1 hold. If we choose $k = cs^{2\alpha/(2\alpha+d)}$ for some constant c, then*

$$r(\|\cdot\|^2) \lesssim s^{-\alpha(1+\beta)/(2\alpha+d)}.$$

On the other hand, if $k = c(s/\kappa^{K-1})^{2\alpha/(2\alpha+K)}$ or $k = cs^{2\alpha/(2\alpha+K)}$ for some constant c, then

$$r(D^*) \lesssim (s/\kappa^{K-1})^{-\alpha(1+\beta)/(2\alpha+K)} \qquad \text{and} \qquad r(D^{**}) \lesssim s^{-\alpha(1+\beta)/(2\alpha+K)}.$$

Let $\mathcal{F}$ be the collection of regression function $\eta(x)$ and probability density function $\mu(x)$ satisfying Assumption 1. We have

$$\min_k \sup_{(\eta,\mu)\in\mathcal{F}} r(\|\cdot\|^2) \gtrsim s^{-\alpha(1+\beta)/(2\alpha+d)},$$

$$\min_k \sup_{(\eta,\mu)\in\mathcal{F}} r(D^*) \gtrsim (s/\kappa^{K-1})^{-\alpha(1+\beta)/(2\alpha+K)} \quad \text{and} \quad \min_k \sup_{(\eta,\mu)\in\mathcal{F}} r(D^{**}) \gtrsim s^{-\alpha(1+\beta)/(2\alpha+K)}.$$

We write $a \lesssim b$ for two sequences a and b if there exists a constant C such that $a \leq Cb$, and $a \gtrsim b$ for two sequences a and b if there exists a constant c such that $a \geq cb$. The two parts in Theorem 2 show that the convergence rates are tight. Theorem 2 suggests that when the target distances D^* and D^{**} are used, the curse of dimensionality is alleviated and the convergence rate of k-NN can be much improved. The reason for the improvement is that the neighborhood defined by target distance D^* and D^{**} can better fit the geometry of the Bayes classification rule than that defined by Euclidean distance. To illustrate this point, we compare balls defined by Euclidean distance and target distance, respectively, denoted by $\mathcal{B}_{\|\cdot\|^2}(x, r)$ and $\mathcal{B}_{D^*}(x, r)$. The shapes of

the two neighborhoods are quite different: $\mathcal{B}_{\|\cdot\|^2}(x, r)$ is a standard sphere, while $\mathcal{B}_{D^*}(x, r)$ is a cylinder, of which axis is in the orthogonal complement of B. One toy example in $\mathbb{R}^2$ is illustrated in Figure 2, where the red area is $\mathcal{B}_{D^*}(x, r)$, and the yellow area is $\mathcal{B}_{\|\cdot\|^2}(x, r)$. As pointed out by Proposition 1, the value of $\eta(x)$ only changes along with the directions in the column subspace of U, so we can expect values of $\eta(x)$ is more similar in $\mathcal{B}_{D^*}(x, r)$ than in $\mathcal{B}_{\|\cdot\|^2}(x, r)$ and thus $\mathcal{B}_{D^*}(x, r)$ can lead to a smaller bias than $\mathcal{B}_{\|\cdot\|^2}(x, r)$.

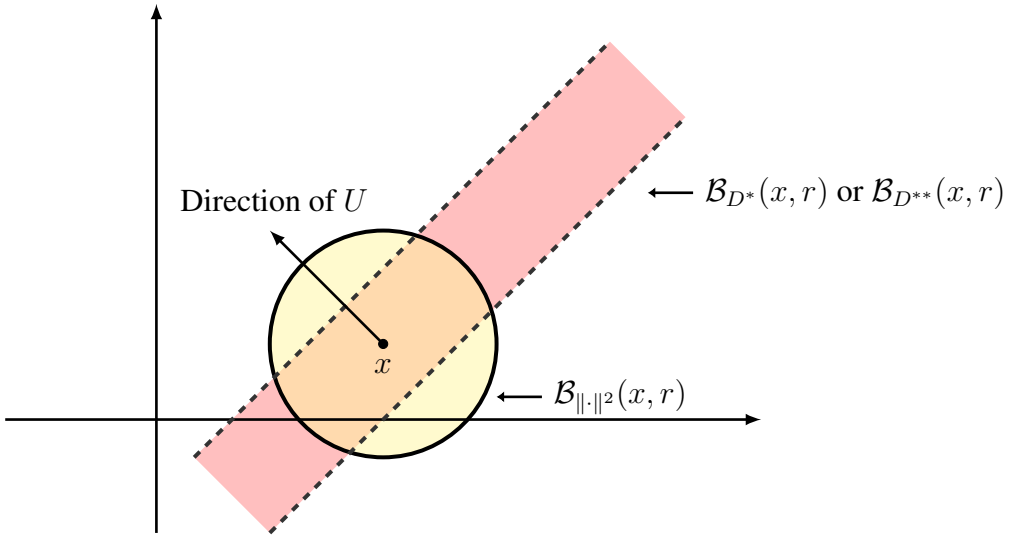

Figure 2: An illustrative example for the neighborhoods defined by Euclidean distance $\|\cdot\|^2$ and target distance D^* or D^{**}.

4.2 Two-Sample Testing

Two-sample testing is central to statistical inferences and an important tool in many applications. Unlike the multi-view data used for metric learning, we observe only one view but with labels for each sample in the standard two-sample testing setting. Specifically, the data we observe in two-sample testing is $(X_1, Y_1), \ldots, (X_s, Y_s)$ and we are interested in the following hypothesis

$$H_0 : \mathbb{E}(X|Y = -1) = \mathbb{E}(X|Y = 1) \qquad \text{and} \qquad H_1 : \mathbb{E}(X|Y = -1) \neq \mathbb{E}(X|Y = 1).$$

In order to test such a hypothesis, many different tests have been proposed. One of the most widely used test families is the distance-based method, including the energy distance test (Székely

and Rizzo, 2005; Sejdinovic et al., 2013), permutational multivariate analysis of variance (PERMANOVA) (McArdle and Anderson, 2001; Anderson, 2014; Wang, Cai, and Li, 2021), and graph-based test (Friedman and Rafsky, 1979; Chen and Friedman, 2017). The idea of a distance-based test is that the pairwise distances between samples are first evaluated, and then the test is then constructed based on the distance matrix. The distance-based two-sample test is also closely related to the kernel-based two-sample test, such as the maximum mean discrepancy (MMD) (Gretton et al., 2012). In particular, Sejdinovic et al. (2013) shows the equivalence between the energy distance test and the MMD test when the distance is a metric of negative type.

In this section, we mainly focus on the energy distance test

$$E(D) = \frac{2}{s_+ s_-} \sum_{Y_i \neq Y_{i'}} D(X_i, X_{i'}) - \frac{1}{s_+(s_+ - 1)} \sum_{Y_i = Y_{i'} = 1} D(X_i, X_{i'}) - \frac{1}{s_-(s_- - 1)} \sum_{Y_i = Y_{i'} = -1} D(X_i, X_{i'}),$$

where D is a given distance, $s_+ = |\{i : Y_i = 1\}|$, and $s_- = |\{i : Y_i = -1\}|$. The energy distance test compares the average within-group distance and the one across groups and can fully characterize the distribution homogeneity between groups when the distance is a metric of negative type (Sejdinovic et al., 2013). Euclidean distance is a metric of negative type, but neither D^* nor D^{**} is since they measure the difference only along with K directions. This suggests that the target distances in self-supervised metric learning cannot fully capture the difference between two general distributions but are particularly suitable for the multi-view data, as we show in this section. To make decisions, we still need to choose a critical value for $E(D)$ or transform $E(D)$ to a p-value. Here, we consider two different ways to make decisions based on $E(D)$. The first one we consider here is the permutation test. Specifically, let Φ_s be the set of permutations on $\{1, \ldots, s\}$, i.e., $\Phi_s = \{\phi : \{1, \ldots, s\} \to \{1, \ldots, s\} | \phi(i) \neq \phi(j) \text{ if } i \neq j\}$. Given a permutation ϕ, we write $\phi E(D)$ as the energy distance test statistic calculated on $(X_1, Y_{\phi(1)}), \ldots, (X_s, Y_{\phi(s)})$. Let $\phi_1, \ldots, \phi_B$ be B permutations drawn from Φ_s randomly. Then, the p-value can be calculated by

$$\hat{P} = \frac{1 + \sum_{b=1}^{B} \mathbf{I}_{(\phi_b E(D) \geq E(D))}}{1 + B}.$$

We reject the null hypothesis when $\hat{P} \leq \alpha$. The second way to make the decision is based on asymptotic distribution. We show that under the null hypothesis, $E(D)/\mathrm{sd}_{H_0}(E(D)) \to N(0,1)$, where $\mathrm{sd}_{H_0}(E(D))$ is the standard deviation of $E(D)$ under the null hypothesis. So we can reject the null hypothesis when $E(D) > z_\alpha \mathrm{sd}_{H_0}(E(D))$ where z_α is the upper α-quantile of standard normal distribution. $\mathrm{sd}_{H_0}(E(D))$ is usually a function of the covariance matrix and thus can be estimated consistently in practice (Chen and Qin, 2010).

The energy distance test's performance depends largely on the choice of distance and the difference between distributions in two groups. Here, we mainly study the tests' performance when the means between groups, $\mu = \mathbb{E}(X|Y=-1) \neq \mathbb{E}(X|Y=1)$, are different. We consider detection radius for the two-sample testing problem to compare the performance of different distances

$$r(D,\epsilon) = \inf\left\{ r : \underbrace{\mathbb{P}(\phi_D = 1|H_0)}_{type\ I\ error} + \underbrace{\mathbb{P}(\phi_D = 0|H_1(r))}_{type\ II\ error} \leq \epsilon \right\},$$

where ϕ_D is the test defined above by permutation test or asymptotic distribution and $H_1(r) = \{\|\mu\| \geq r\}$. Intuitively, the detection radius $r(D,\epsilon)$ represents the smallest distance to separate the null and alternative hypothesis reliably. Thus, the test is more powerful to distinguish similar samples when $r(D,\epsilon)$ is smaller. To characterize the performance of energy distance test, we make the following assumptions.

Assumption 2. *It holds that*

(a) we choose $\alpha = \epsilon/2$*;*

(b) assume $\mathbb{P}(Y=1) = \mathbb{P}(Y=-1) = 1/2$*;*

(c) assume the covariance matrix of $\epsilon_{i,j}$ *is* Σ

(d) if we write the covariance matrix of X *given* $Y=1$ *as* Σ_+ *and the covariance matrix of* X *given* $Y=-1$ *as* Σ_-*, then we assume* $\mathrm{Tr}(\Sigma_{i_1}\Sigma_{i_2}\Sigma_{i_3}\Sigma_{i_4}) = o\left(\|\Sigma_+ + \Sigma_-\|_F^4\right)$ *for* $i_1, i_2, i_3, i_4 = +$ *or* $-$*. We assume it still hold when we replace* Σ_+ *and* Σ_- *by* $B^T\Sigma_+B$ *and* $B^T\Sigma_-B$ *(*$U^T\Sigma_+U$ *and* $U^T\Sigma_-U$*).*

(e) for any $1 \le i < j \le s$*, we assume* $\mathbb{E}(X_i^T X_j)^4 = o\left(s\|\Sigma_+ + \Sigma_-\|_F^4\right)$*,* $\mathbb{E}(X_i^T BB^T X_j)^4 = o\left(s\|B^T(\Sigma_+ + \Sigma_-)B\|_F^4\right)$ *and* $\mathbb{E}(X_i^T UU^T X_j)^4 = o\left(s\|U^T(\Sigma_+ + \Sigma_-)U\|_F^4\right)$.

The first three assumptions in Assumption 2 are fairly weak conditions, and the last two are moment conditions used for the central limit theorem of U-statistics. Similar assumptions also appear in Hall (1984); Chen and Qin (2010); Li and Yuan (2019). If we use Euclidean distance and the distance D^* and D^{**} in the energy distance test $E(D)$, the detection radius can be characterized by the following theorem.

Theorem 3. *Suppose assumptions in Section 2 and Assumption 2 hold. If the test* ϕ_D *is defined by permutation test (permutation test does not need (d) and (e) in Assumption 2) or asymptotic distribution, then*

$$r(\|\cdot\|^2, \epsilon) \lesssim \frac{\|BB^T + \Sigma\|_F^{1/2}}{\sqrt{s}}, \qquad r(D^*, \epsilon) \lesssim \frac{\|\Lambda^2 + B^T\Sigma B\|_F^{1/2}}{\sqrt{s\lambda_K}}$$

and

$$r(D^{**}, \epsilon) \lesssim \frac{\|\Lambda + U^T\Sigma U\|_F^{1/2}}{\sqrt{s}}.$$

Consider the energy distance test defined by permutation test or asymptotic distribution and the following local alternative hypothesis $\tilde{H}_1(r) = \{\mu = ru_K\}$*. If* $r = o(\|BB^T + \Sigma\|_F^{1/2}/\sqrt{s})$*, then*

$$\mathbb{P}(\phi_{\|\cdot\|^2} = 0 | \tilde{H}_1(r)) \to 1 - \alpha.$$

Similarly, if $r = o(\|\Lambda^2 + B^T\Sigma B\|_F^{1/2}/\sqrt{s\lambda_K})$ *or* $r = o(\|\Lambda + U^T\Sigma U\|_F^{1/2}/\sqrt{s})$*, then*

$$\mathbb{P}(\phi_{D^*} = 0 | \tilde{H}_1(r)) \to 1 - \alpha \qquad \text{and} \qquad \mathbb{P}(\phi_{D^{**}} = 0 | \tilde{H}_1(r)) \to 1 - \alpha.$$

Together with the first and second part of Theorem 3, the detection radius for Euclidean distance and the target distances of self-supervised metric learning are sharp. Theorem 3 suggests that the detection radius of the energy distance test is mainly determined by the variation of X, which can be decomposed into two parts: the first part corresponds to the difference between samples and the second part is due to the variation between different views of the same sample. If we assume

$\Sigma = \sigma^2 I$ in Theorem 3, we can have

$$r(\|\cdot\|^2, \epsilon) \lesssim \frac{(\sqrt{K}\lambda_1 + \sqrt{d}\sigma^2)^{1/2}}{\sqrt{s}}, \qquad r(D^*, \epsilon) \lesssim \sqrt{\kappa}\frac{(\sqrt{K}\lambda_1 + \sqrt{K}\sigma^2)^{1/2}}{\sqrt{s}}$$

and

$$r(D^{**}, \epsilon) \lesssim \frac{(\sqrt{K}\lambda_1 + \sqrt{K}\sigma^2)^{1/2}}{\sqrt{s}}.$$

When self-supervised metric learning is used, variation between different views can be reduced from $\sqrt{d}\sigma^2$ to $\sqrt{K}\sigma^2$. It implies that the energy distance test can be improved by self-supervised metric learning when the variation between different views dominates, i.e., $\sqrt{K}\lambda_1 \ll \sqrt{d}\sigma^2$.

5 Self-Supervised Metric Learning in Multi-View Data

5.1 Data-Driven Distance on Downstream Tasks

In the previous section, we show that target distances D^* and D^{**} in self-supervised metric learning are good distances for downstream analysis. However, we cannot directly adopt target distances in each downstream task as they are usually unknown in advance. In practice, we still need to estimate D^* and D^{**} from the unlabeled multi-view data. One may wonder if the data-driven distances estimated from unlabeled multi-view data can also improve the downstream tasks similarly to target distances. Our investigation in this section confirms that the data-driven distance can benefit the downstream analysis when the target distances can be estimated accurately. It is sufficient to estimate the following matrices to estimate the target distances

$$M^* = BB^T \qquad \text{and} \qquad M^{**} = UU^T.$$

Let $\hat{M}^*$ and $\hat{M}^{**}$ be some estimators for M^* and M^{**}, and $D_{\hat{M}^*}$ and $D_{\hat{M}^{**}}$ be the distances defined by them. The measure $\Delta(D, D')$ can be rewritten as the spectral norm of matrix difference, $\Delta(D, D') = \|M - M'\|$, where $D(X_1, X_2) = (X_1 - X_2)^T M (X_1 - X_2)$ and $D'(X_1, X_2) = (X_1 - X_2)^T M' (X_1 - X_2)$. The following theorem shows that the estimated distances can still improve downstream analysis.

Theorem 4. *Suppose the data in self-supervised metric learning is independent from the data in downstream tasks and assumptions in Section 2 hold and κ is bounded. Let $\hat{M}^*$ and $\hat{M}^{**}$ be some estimators of M^* and M^{**} such that*

$$\Delta(D^*, D_{\hat{M}^*}) \leq \delta^* \qquad \text{and} \qquad \Delta(D^{**}, D_{\hat{M}^{**}}) \leq \delta^{**}.$$

- *(k-nearest neighbor classification) Suppose Assumption 1 holds and let c be some constant. If $k = c(s/\kappa^{K-1})^{2\alpha/(2\alpha+K)}$, $\delta^* \lesssim \lambda_K(s/\kappa^{K-1})^{-1/(2\alpha+K)}$ in $D_{\hat{M}^*}$ or $k = cs^{2\alpha/(2\alpha+K)}$, $\delta^{**} \lesssim s^{-1/(2\alpha+K)}$ in $D_{\hat{M}^{**}}$, then*

$$r(D_{\hat{M}^*}) \lesssim (s/\kappa^{K-1})^{-\alpha(1+\beta)/(2\alpha+K)} \qquad \text{and} \qquad r(D_{\hat{M}^{**}}) \lesssim s^{-\alpha(1+\beta)/(2\alpha+K)}.$$

- *(two-sample testing) Suppose Assumption 2 and $\|\Sigma\| \lesssim \lambda_1$ hold and let c be a large enough constant. If $\delta^* = o(\lambda_K)$ in $D_{\hat{M}^*}$ or $\delta^{**} = o(1)$ in $D_{\hat{M}^{**}}$, then*

$$r(D_{\hat{M}^*}, \epsilon) \lesssim \frac{\|\Lambda^2 + B^T\Sigma B\|_F^{1/2}}{\sqrt{s}} \qquad \text{and} \qquad r(D_{\hat{M}^{**}}, \epsilon) \lesssim \frac{\|\Lambda + U^T\Sigma U\|_F^{1/2}}{\sqrt{s}}.$$

- *(k-means clustering) Suppose Assumption S1 holds and $t > \log s$. If $\|B^T\mu\| \gg \Psi(\Lambda^2 + B^T\Sigma_{\pm}B)$, $\delta^* = o(\lambda_K)$ in $D_{\hat{M}^*}$ or $\|\mu\| \gg \Psi(\Lambda + U^T\Sigma_{\pm}U)$, $\delta^{**} = o(1)$ in $D_{\hat{M}^{**}}$, then*

$$r(D_{\hat{M}^*}) \leq \Gamma(1 + o(1), B^T\mu, B^T\Sigma_{\pm}B) \qquad \text{and} \qquad r(D_{\hat{M}^{**}}) \leq \Gamma(1 + o(1), \mu, U^T\Sigma_{\pm}U)$$

  with probability at least $1 - s^5 - \exp(-\sqrt{v}\|\mu\|)$ where $v \to \infty$.

- *(sample identification) Suppose Assumption S2 holds and $\lambda_d(\Sigma) \geq c\|\Sigma\|$ where $\lambda_d(\Sigma)$ is the smallest eigenvalue of Σ. If $\delta^* = o(\lambda_K)$ in $D_{\hat{M}^*}$ or $\delta^{**} = o(1)$ in $D_{\hat{M}^{**}}$, then*

$$r(D_{\hat{M}^*}, \epsilon) \lesssim \frac{\|B^T\Sigma B\|_F^{1/2}}{\lambda_K} \quad \text{and} \quad r(D_{\hat{M}^{**}}, \epsilon) \lesssim \frac{\|U^T\Sigma U\|_F^{1/2}}{\sqrt{\lambda_K}}.$$

Theorem 4 suggests that the estimated distance $D_{\hat{M}^*}$ and $D_{\hat{M}^{**}}$ from the self-supervised metric learning could help achieve a similar performance as D^* and D^{**} when the target distances can be estimated accurately. Self-supervised learning can help improve two-sample testing, k-means clustering, and sample identification as long as we have enough unlabeled multi-view data to estimate the target distance consistently, i.e., $\Delta(D^*, D_{\hat{M}^*}) = o(\lambda_K)$ or $\Delta(D^{**}, D_{\hat{M}^{**}}) = o(1)$. Unlike

these three downstream tasks, the improvement of k-nearest neighbor classification needs a more accurate estimation of target distance. Theorem 4 assumes the independence between data in metric learning and downstream tasks for the simplicity of analysis. This is a reasonable assumption when we have many unlabeled multi-view data in a typical self-supervised learning setting. If the metric learning and downstream tasks use the same data set, the results in Theorem 4 might still hold, but the analysis can be much more involved.

5.2 Spectral Self-Supervised Metric Learning

The previous section shows that the downstream task can be improved when the target distances can be estimated accurately. Two questions naturally arise: how shall we estimate the target distances? how much unlabeled multi-view data is sufficient to improve the downstream analysis? To answer these questions, we consider a spectral method to estimate D^* and D^{**} in this section. Since M^* is the optimal solution of (3), a natural idea of estimating M^* is to replace $\mathbb{E}\left(D_M(X_{i,j}, X_{i',j'}) - D_M(X_{i,j}, X_{i,j'})\right)$ with its empirical version. More concretely, its empirical version can be written as

$$\frac{1}{m(m-1)n^2} \sum_{i \neq i', j, j'} D_M(X_{i,j}, X_{i',j'}) - \frac{1}{mn(n-1)} \sum_{i, j \neq j'} D_M(X_{i,j}, X_{i,j'}).$$

Here, we consider all pairs of dissimilar and similar data and use U-statistics as the estimator. After plugging in the empirical version of distance difference and some calculation, M^* can be estimated by the following optimization problem

$$\max_M \operatorname{Tr}\left(\hat{R}M\right), \qquad \text{s.t.} \quad \|M\|_F \leq 1 \quad \text{and} \quad \operatorname{rank}(M) \leq K.$$

where $\hat{R}$ is a $d \times d$ matrix

$$\hat{R} = \frac{1}{mn(n-1)} \sum_{i, j \neq j'} \left(X_{i,j} X_{i,j'}^T + X_{i,j'} X_{i,j}^T\right) - \frac{1}{m(m-1)} \sum_{i \neq i'} \left(\bar{X}_i \bar{X}_{i'}^T + \bar{X}_{i'} \bar{X}_i^T\right)$$

Here, $\hat{R}$ is an unbiased estimator of BB^T regardless of the $\epsilon_{i,j}$'s distribution. The reason for having unbiased estimator is that we observe several views of each sample. This is different from the classical factor model, where we only observe a single view for each sample (Fan et al., 2020).

In the above optimization problem, we also add a constraint for the rank of M since BB^T is a low-rank matrix. This optimization problem's form can then naturally lead to a simple spectral algorithm to estimate M^*, summarized in Algorithm 1. The spectral method in can also be easily adjusted to estimate M^{**} when we change the last step, which is also included in Algorithm 1.

Algorithm 1 Spectral Metric Learning in Multi-view Data

Input: Multi-view data $(X_{i,1}, \ldots, X_{i,n_i})$ for $i = 1, \ldots m$.
Output: A matrix $\hat{M}^*$ or $\hat{M}^{**}$.
Evaluate $\hat{R}$.
Find the first K eigenvalues and eigenvectors of $\hat{R}$, i.e., $(\hat{\lambda}_1, \ldots, \hat{\lambda}_K)$ and $(\hat{u}_1, \ldots, \hat{u}_K)$.
Estimate $\hat{M}^*$ or $\hat{M}^{**}$ by

$$\hat{M}^* = \sum_{k=1}^{K} \hat{\lambda}_k \hat{u}_k \hat{u}_k^T \qquad \text{or} \qquad \hat{M}^{**} = \sum_{k=1}^{K} \hat{u}_k \hat{u}_k^T.$$

The Algorithm 1 seems computationally expensive at first sight since the definition of $\hat{R}$ involves U-statistics, which usually requires quadratic time complexity. However, thanks to the special structure of empirical covariance matrix $\hat{R}$, it can be rewritten as the following equivalent form

$$\hat{R} = \left(\frac{n}{n-1} + \frac{1}{m-1} \right) \frac{1}{m} \sum_i \bar{X}_i \bar{X}_i^T - \frac{1}{mn(n-1)} \sum_{i,j} X_{i,j} X_{i,j}^T - \frac{m}{m-1} \bar{\bar{X}} \bar{\bar{X}}^T$$

where $\bar{X}_i = n^{-1} \sum_j X_{i,j}$ and $\bar{\bar{X}} = m^{-1} \sum_i \bar{X}_i$. Thus, $\hat{R}$ can be computed in a linear time.

We now investigate the theoretical properties of $\hat{M}^*$ or $\hat{M}^{**}$ in Algorithm 1. To the end, we make the following assumptions.

Assumption 3. *It holds that*

(a) $\epsilon_{i,j}$ and Z_i follow sub-Gaussian distributions, that is, for any $a \in \mathbb{R}^d$ and $b \in \mathbb{R}^K$

$$\mathbb{E}\left(e^{\langle a, \epsilon_{i,j} \rangle}\right) \leq e^{\sigma^2 \|a\|^2/2} \qquad \text{and} \qquad \mathbb{E}\left(e^{\langle b, Z_i - \mathbb{E}(Z_i) \rangle}\right) \leq e^{\|b\|^2/2};$$

(b) conditional number κ is bounded;

(c) assume K is known.

The assumption on sub-Gaussian distributions is the key assumption in Assumption 3, which is commonly used in the study of eigenspace estimation (Zhang, Cai, and Wu, 2018; Chen et al., 2020b). Since we observe multi-view data of each sample, we do not assume diagonal or sparse covariance matrix as literature (Yao, Zheng, and Bai, 2015; Zhang, Cai, and Wu, 2018). The following theorem characterizes the convergence rate of $\hat{M}^*$ and $\hat{M}^{**}$.

Theorem 5. *Suppose assumptions in Section 2 and Assumption 3 hold. If* $m \geq c\log(d+m)(\kappa^2 K\log(d+m) + d\kappa\sigma^2/n\lambda_K + d\sigma^4/n^2\lambda_K^2)$ *for a large enough constant* c*, then, with probability at least* $1-6/(d+m)^5$*, we have*

$$\Delta(D^*, D_{\hat{M}^*}) \lesssim \frac{\sqrt{\log(d+m)}}{\sqrt{m}}\left[\sqrt{K}\lambda_1 + \sigma\frac{\sqrt{d\lambda_1}}{\sqrt{n}} + \sigma^2\frac{\sqrt{d}}{n}\right].$$

In addition, if $m \geq c\log(d+m)(K + d\kappa\sigma^2/n\lambda_K + d\sigma^4/n^2\lambda_K^2)$ *for a large enough constant* c*, we have similar results for* $\hat{M}^{**}$*, that is*

$$\Delta(D^{**}, D_{\hat{M}^{**}}) \lesssim \frac{\sqrt{\log(d+m)}}{\sqrt{m}}\left[\sigma\frac{\sqrt{d}}{\sqrt{n\lambda_K}} + \sigma^2\frac{\sqrt{d}}{n\lambda_K}\right]$$

with probability at least $1-6/(d+m)^5$.

Naturally, one may wonder whether the bound for spectral method is tight, and if there are some other methods that can help learn distance D^* or D^{**} better. To answer these questions, we develop the information-theoretic lower bound that matches the upper bound in Theorem 5. To develop the lower bound, we focus on the following Gaussian noise model $X_{i,j} = BZ_i + \epsilon_{i,j}$, where $Z_i \sim N(0, I)$ and $\epsilon_{i,j} \sim N(0, \sigma^2 I)$ and consider the collection of matrix B

$$\mathcal{B}(\nu) = \left\{B \in \mathbb{R}^{d\times K} : \lambda_1(B)/\lambda_K(B) \leq \sqrt{\kappa}, \lambda_K(B) \geq \sqrt{\nu}\right\},$$

where $\lambda_1(B)$ and $\lambda_K(B)$ are the largest and smallest singular value of B.

Theorem 6. *Suppose* $\kappa > 1$ *is bounded,* $m > K$ *and* $4K \leq d$*. Then*

$$\inf_{\hat{M}^*}\sup_{B\in\mathcal{B}(\nu)} \mathbb{E}\left(\Delta(D^*, D_{\hat{M}^*})\right) \gtrsim \frac{1}{\sqrt{m}}\left[\sqrt{K}\nu + \sigma\frac{\sqrt{d\nu}}{\sqrt{n}} + \sigma^2\frac{\sqrt{d}}{n}\right].$$

We also have similar results for $\hat{M}^{**}$, that is

$$\inf_{\hat{M}^{**}} \sup_{B\in\mathcal{B}(\nu)} \mathbb{E}\left(\Delta(D^{**}, D_{\hat{M}^{**}})\right) \gtrsim \frac{1}{\sqrt{m}}\left[\sigma\frac{\sqrt{d}}{\sqrt{n\nu}} + \sigma^2\frac{\sqrt{d}}{n\nu}\right].$$

Through comparing Theorem 5 and 6, we can know the results in Theorem 5 are indeed sharp up to a logarithm factor. As shown in these two theorems, estimating $D_{\hat{M}^{**}}$ is easier than $D_{\hat{M}^{*}}$ since there is no need for estimating the eigenvalues $\hat{\lambda}_1, \ldots, \hat{\lambda}_K$. Theorem 5 also suggests $D_{\hat{M}^{*}}$ and $D_{\hat{M}^{**}}$ can improve the downstream analysis provided the sample size of unlabeled multi-view data is large enough. By combining Theorem 4 and 5, we have the following corollary which precisely characterizes the sample size needed for downstream tasks improvement.

Corollary 1. *Suppose assumptions in Theorem 4 and 5 hold. If $m \gg (s/\kappa^{K-1})^{1/(2\alpha+K)}\log(d+m)(K + d\sigma^2/n\lambda_K + d\sigma^4/n^2\lambda_K^2)$ in $D_{\hat{M}^{*}}$ or $m \gg s^{1/(2\alpha+K)}\log(d+m)(d\sigma^2/n\lambda_K + d\sigma^4/n^2\lambda_K^2)$ in $D_{\hat{M}^{**}}$, k-NN can achieve the same convergence rate in Theorem 4. If $m \gg \log(d+m)(K + d\sigma^2/n\lambda_K + d\sigma^4/n^2\lambda_K^2)$ in $D_{\hat{M}^{*}}$ or $m \gg \log(d+m)(d\sigma^2/n\lambda_K + d\sigma^4/n^2\lambda_K^2)$ in $D_{\hat{M}^{**}}$, two-sample testing, k-means clustering and sample identification can also achieve the same convergence rate in Theorem 4.*

6 Numerical Experiments

In this section, we conduct several numerical experiments to complement our theoretical developments. In particular, we compare the performance of the four downstream tasks in Section 4 when Euclidean distance and resulting distance from metric learning are used.

6.1 Simulated Data

To simulate the data, we consider the Gaussian model $X_{i,j} = BZ_i + \epsilon_{i,j}$, where $\epsilon_{i,j} \sim N(0, \sigma^2 I)$. Here, we choose $\lambda_k = \lambda(K-k+1)/K$ for some λ and the directions of B, $\{b_1/\|b_1\|, \ldots, b_K/\|b_K\|\}$, are obtained from the first K left-singular vectors of randomly generated $d \times d$ standard Gaussian matrix. We generate Z_i from a mixture model, $0.5N(\alpha, I - \alpha\alpha^T) + 0.5N(-\alpha, I - \alpha\alpha^T)$, for some $\alpha \in \mathbb{R}^K$ with $\|\alpha\| < 1$. We let $Y_i = 1$ if Z_i is drawn from $N(\alpha, I - \alpha\alpha^T)$ and $Y_i = -1$ otherwise.

Sample identification To study the effect of $\|Z_1 - Z_2\|$ and K, we vary $\|Z_1 - Z_2\| = 1, 2, 3, 4, 5$ and $K = 10, 50$. Specifically, we set the first $K/2$ elements in $Z_1 - Z_2$ as zero and the last $K/2$ elements in $Z_1 - Z_2$ as the same non-zero constant. We consider 7 distances: Euclidean distance, target distance D^* and D^{**}, estimated distance D^* and D^{**} by spectral method with $m = 1000, 5000$ samples. We choose $d = 100$, $\lambda = 4$, $\sigma^2 = 1$ and $n = 10$ and repeat the simulation 500 times. We compare the performance of sample identification by power, which is estimated by the number of rejecting null hypothesis. The results are summarized in Table 3. Table 3 suggests that self-supervised metric learning is indeed helpful for sample identification, and the helps shrinkage when K becomes larger, which is consistent with the theoretical results.

	$K = 10$					$K = 50$				
$\|Z_1 - Z_2\|$	1	2	3	4	5	1	2	3	4	5
$\|\cdot\|^2$	0.08	0.21	0.42	0.77	0.96	0.07	0.13	0.34	0.61	0.91
$\hat{D}^*$ (1000)	0.04	0.24	0.64	0.95	1.00	0.08	0.14	0.28	0.53	0.85
$\hat{D}^*$ (5000)	0.05	0.27	0.65	0.96	1.00	0.08	0.13	0.30	0.56	0.87
D^*	0.06	0.27	0.67	0.97	1.00	0.08	0.13	0.30	0.56	0.87
$\hat{D}^{**}$ (1000)	0.09	0.47	0.90	0.99	1.00	0.09	0.22	0.50	0.83	0.99
$\hat{D}^{**}$ (5000)	0.09	0.48	0.90	1.00	1.00	0.08	0.20	0.49	0.82	0.99
D^{**}	0.09	0.47	0.90	1.00	1.00	0.08	0.20	0.50	0.82	0.99

Table 3: Comparisons of different distances on sample identification.

Two-sample testing We now move to the simulation experiment for two-sample testing. Similar to sample identification, we still compare the same 7 distances and choose $d = 100$, $\sigma^2 = 1$, $K = 10$, $n = 10$ and $s = 500$. Let α be a vector such that $\alpha_1 = \ldots = \alpha_4 = 0$ and $\alpha_5 = \ldots = \alpha_{10} = r/\sqrt{6}$ for some r. We study the effect of $\|\mu\|$ and λ by considering the following two experiment settings: 1) $\lambda = 1$ and $r = 0, 0.05, \ldots, 0.5$ 2) $\lambda = 0.5, 1, \ldots, 5$ and $r = 0.3/\lambda$ so that $\|\mu\|$ is fixed. To evaluate the power of different methods, we still repeat the simulation 500 times. The results are summarized in Figure 3. Through Figure 3, we can conclude that self-supervised metric learning is helpful when λ/σ^2 is moderate, while all distances perform similarly when λ/σ^2 is large. These results help verify the theoretical conclusion in Theorem 3.

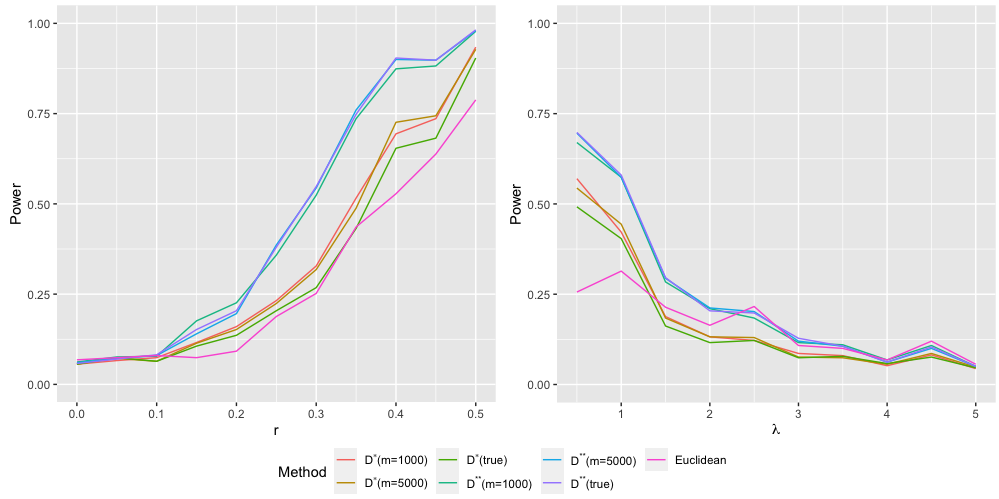

Figure 3: Comparisons of different distances on two-sample testing: left is setting 1 and right is setting 2.

k-means clustering We then consider the simulation experiment for k-means clustering. We adopt the same setting in two-sample testing and set $\lambda = 2$. We choose α as a vector such that $\alpha_1 = \ldots = \alpha_4 = r/\sqrt{4}$ for some r and $\alpha_5 = \ldots = \alpha_{10} = 0$. To compare the required signal, we vary $r = 0.4, 0.6, 0.8, 1$ and use the mis-clustering rate as the measure of performance, which is defined in Section S1.1. We consider two ways to choose the initial estimator of mean in k-means: 1) we randomly choose the two points as initial points 2) we use the true mean in each class as initial points. The results based on the 500 times simulation are summarized in Table 4. In Table 4, even when the starting point is perfect, the performances of D^* is not as good as Euclidean distance and D^{**} due to the anisotropic transformation. Moreover, the distance D^{**} is slightly helpful when random initial points are used. This is again consistent with the theoretical results.

k-nearest neighbor classification In the last simulation experiment, we compare the performance of k-nearest neighbor classification when it works with different distances. We use the same setting in k-means clustering and vary s and r in α, where $\alpha_1 = \ldots = \alpha_5 = 0$ and $\alpha_6 = \ldots = \alpha_{10} = r/\sqrt{5}$. Specifically, we consider the following two experiment settings: $r = 0.9$ and the sample size is different $s = 500, 1000, \ldots, 5000$; sample size is $s = 2000$ and $r = 0.1, \ldots, 1$. The misclassification error defined in Section 4.1 is used as the measure for perfor-

	Random Start				Perfect Start			
	$r = 0.4$	$r = 0.6$	$r = 0.8$	$r = 1$	$r = 0.4$	$r = 0.6$	$r = 0.8$	$r = 1$
$\|\cdot\|^2$	0.43	0.39	0.34	0.14	0.38	0.31	0.21	0.05
$\hat{D}^*$ (1000)	0.43	0.40	0.36	0.23	0.41	0.37	0.29	0.12
$\hat{D}^*$ (5000)	0.43	0.39	0.34	0.23	0.41	0.37	0.31	0.14
D^*	0.43	0.39	0.34	0.24	0.41	0.37	0.31	0.15
$\hat{D}^{**}$ (1000)	0.43	0.39	0.34	0.12	0.40	0.34	0.24	0.05
$\hat{D}^{**}$ (5000)	0.43	0.39	0.34	0.12	0.39	0.34	0.24	0.05
D^{**}	0.43	0.39	0.34	0.13	0.40	0.34	0.24	0.05

Table 4: Comparisons of different distances on k-means clustering.

mance of different distances. The results are summarized in Figure 4, showing the self-supervised metric learning is helpful for k-NN, and the error decreases when the sample size or the difference between populations increases (large r implies large β in marginal assumption).

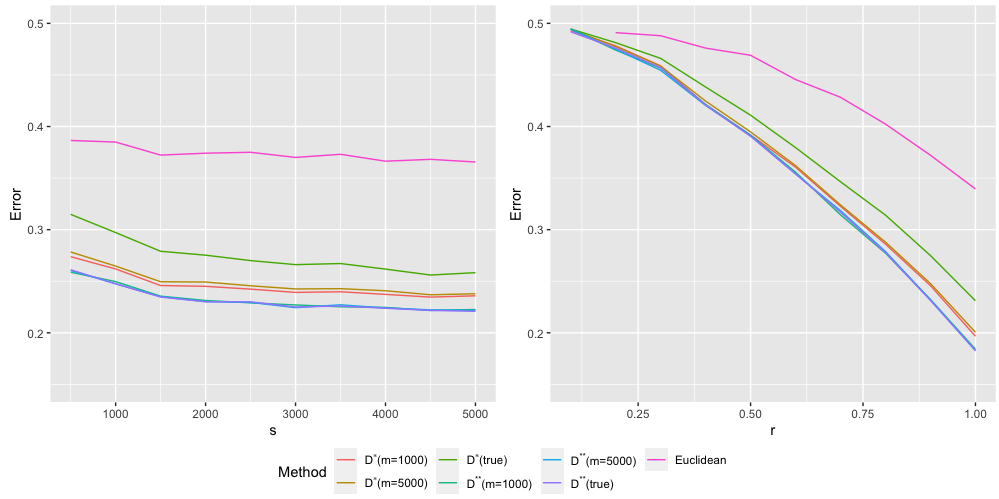

Figure 4: Comparisons of different distances on k-nearest neighbor classification.

All the numerical results in these four simulation experiments are consistent with theoretical conclusion in Section 4. Compared with target distance D^*, the isotropic target distance D^{**} is a better choice for all four downstream tasks we consider here. In addition, distance estimated from self-supervised metric learning performs almost as well as the true target distance in these simulation experiments.

6.2 Computer Vision Task

We further compare Euclidean distance and resulting distance from self-supervised metric learning on some computer vision tasks. Specifically, we consider two datasets: MNIST (LeCun et al., 1998) and Fashion-MNIST (Xiao, Rasul, and Vollgraf, 2017). Both datasets contain 6×10^4 training images and 10^4 testing images, which are all 28×28 gray-scale images from 10 classes. The difference between the two datasets is that MNIST is a collection of handwritten digits while Fashion-MNIST is a collection of clothing. MNIST and Fashion-MNIST do not contain multi-view data, but we can generate a multi-view dataset by shifting the images. Specifically, we shift the image in 4 different directions (left, right, upper and lower) to generate the multi-view dataset. A toy example of image shifting can be found in Figure 5.

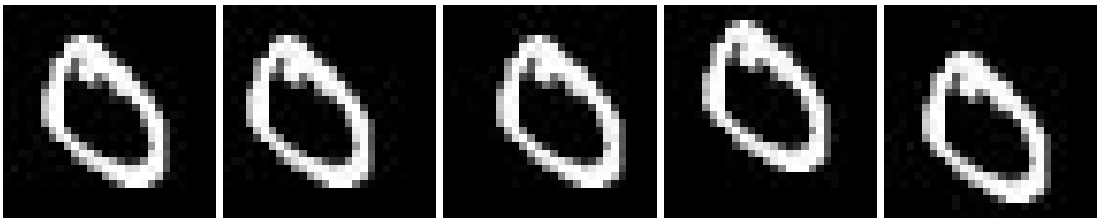

Figure 5: Multi-view data generated from MNIST dataset: from left to right are original, left shift, right shift, upper shift and lower shift.

In each dataset, we consider applying k-NN to classify the images. In this numerical experiment, a large unlabeled multi-view dataset ($m = 10^4$ and $n = 5$) and a small labeled dataset ($s = 10^3, 2 \times 10^3, 5 \times 10^3$) are randomly drawn from training images and then used to train a k-NN classifier. We consider the following three ways to train k-NN classifier: 1) Euclidean distance is used to train k-NN directly on the small labeled dataset; 2) the anisotropic distance D^* is estimated by the spectral method from the unlabeled multi-view dataset, and then the estimated distance is used to train k-NN; 3) the isotropic distance D^{**} is estimated from the unlabeled multi-view dataset and then used to train k-NN. To measure the performances, we adopt the misclassification errors, which can be estimated on 10^3 images randomly drawn from testing images. The misclassification errors are reported in Table 5. It suggests that the self-supervised metric learning on the dataset from simple image shifting is helpful for the downstream classification task.

	MNIST			Fashion-MNIST		
	$\|\cdot\|^2$	D^*	D^{**}	$\|\cdot\|^2$	D^*	D^{**}
$s = 1000$	0.115	0.268	0.094	0.254	0.380	0.254
$s = 2000$	0.086	0.222	0.079	0.240	0.352	0.233
$s = 5000$	0.062	0.169	0.059	0.208	0.318	0.204

Table 5: Comparisons of different distances on computer vision task.

7 Conclusion

This paper conducts a systematic investigation of self-supervised metric learning in unlabeled multi-view data from a downstream task perspective. Building on a latent factor model for multi-view data, we provide theoretical justification for the success of this popular approach. Our analysis precisely characterizes the improvement by self-supervised metric learning on several downstream tasks, including sample identification, two-sample testing, k-means clustering, and k-nearest neighbor classification. Furthermore, we also establish the upper bound on distance estimation's accuracy and the number of samples sufficient for downstream task improvement. We assume that the number of factors K is known in the analysis. In practice, some data-driven methods can help choose K, like Kaiser criterion and scree plot, when it is unknown. See more discussion in Chapter 10 of Fan et al. (2020). The results in this paper rely on the assumption of the latent factor model and are designed for Mahalanobis distance. It could also be interesting to explore if the results can be extended to the deep neural network-based metric learning methods.

ACKNOWLEDGMENT

We thank the editor, associate editor and referees for valuable suggestions. This project is supported by the grant from the National Science Foundation (DMS-2113458).

SUPPLEMENTARY MATERIALS

We provide some extra results and prove all the theorems and relevant lemmas in the online Supplementary Materials. All analyses for numerical experiments can be found under `https://github.com/lakerwsl/SSTMetric-Manuscript-Code`.

References

Adragni, K. P., and Cook, R. D. (2009), "Sufficient dimension reduction and prediction in regression," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367, 4385–4405.

Altman, N. S. (1992), "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, 46, 175–185.

Anderson, M. J. (2014), "Permutational multivariate analysis of variance (PERMANOVA)," *Wiley statsref: statistics reference online*, 1–15.

Arora, S., Khandeparkar, H., Khodak, M., Plevrakis, O., and Saunshi, N. (2019), "A theoretical analysis of contrastive unsupervised representation learning," *arXiv preprint arXiv:1902.09229*.

Audibert, J., and Tsybakov, A. B. (2007), "Fast learning rates for plug-in classifiers," *The Annals of statistics*, 35, 608–633.

Bellet, A., Habrard, A., and Sebban, M. (2013), "A survey on metric learning for feature vectors and structured data," *arXiv preprint arXiv:1306.6709*.

——— (2015), "Metric learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9, 1–151.

Bengio, Y., Courville, A., and Vincent, P. (2013), "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, 35, 1798–1828.

Biau, G., and Devroye, L. (2015), *Lectures on the nearest neighbor method*, vol. 246, Springer.

Cao, Q., Guo, Z., and Ying, Y. (2016), "Generalization bounds for metric and similarity learning," *Machine Learning*, 102, 115–132.

Chechik, G., Sharma, V., Shalit, U., and Bengio, S. (2010), "Large Scale Online Learning of Image Similarity Through Ranking," *Journal of Machine Learning Research*, 11, 1109–1135.

Chen, H., and Friedman, J. H. (2017), "A new graph-based two-sample test for multivariate and object data," *Journal of the American statistical association*, 112, 397–409.

Chen, S., and Qin, Y. (2010), "A two-sample test for high-dimensional data with applications to gene-set testing," *The Annals of Statistics*, 38, 808–835.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020a), "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, PMLR, pp. 1597–1607.

Chen, Y., Chi, Y., Fan, J., and Ma, C. (2020b), "Spectral methods for data science: A statistical perspective," *arXiv preprint arXiv:2012.08496*.

Chopra, S., Hadsell, R., and LeCun, Y. (2005), "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, IEEE, vol. 1, pp. 539–546.

Cvitkovic, M., and Koliander, G. (2019), "Minimal achievable sufficient statistic learning," in *International Conference on Machine Learning*, PMLR, pp. 1465–1474.

Deng, Y., Yuan, Y., Fu, H., and Qu, A. (2021), "Query-augmented Active Metric Learning," *Journal of the American Statistical Association*, 1–36.

Duan, Y., Zheng, W., Lin, X., Lu, J., and Zhou, J. (2018), "Deep adversarial metric learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2780–2789.

Dwibedi, D., Tompson, J., Lynch, C., and Sermanet, P. (2018), "Learning actionable representations from visual observations," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, pp. 1577–1584.

Fan, J., Li, R., Zhang, C., and Zou, H. (2020), *Statistical foundations of data science*, CRC press.

Fearnhead, P., and Prangle, D. (2012), "Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74, 419–474.

Fix, E. (1985), *Discriminatory analysis: nonparametric discrimination, consistency properties*, vol. 1, USAF school of Aviation Medicine.

Flores, G. E., Caporaso, J. G., Henley, J. B., Rideout, J., Domogala, D., Chase, J., Leff, J. W., Vázquez-Baeza, Y., Gonzalez, A., Knight, R., Dunn, R. R., and Fierer, N. (2014), "Temporal variability is a personalized feature of the human microbiome," *Genome biology*, 15, 1–13.

Friedman, J. H., and Rafsky, L. C. (1979), "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *The Annals of Statistics*, 697–717.

Gajer, P., Brotman, R. M., Bai, G., Sakamoto, J., Schütte, U. M., Zhong, X., Koenig, S. S., Fu, L., Ma, Z., Zhou, X., Abdo, Z., Forney, L. J., and Ravel, J. (2012), "Temporal dynamics of the human vaginal microbiota," *Science translational medicine*, 4, 132ra52–132ra52.

Georghiades, A. S., Belhumeur, P. N., and Kriegman, D. J. (2001), "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE transactions on pattern analysis and machine intelligence*, 23, 643–660.

Gidaris, S., Singh, P., and Komodakis, N. (2018), "Unsupervised representation learning by predicting image rotations," *arXiv preprint arXiv:1803.07728*.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012), "A kernel two-sample test," *The Journal of Machine Learning Research*, 13, 723–773.

Guillaumin, M., Verbeek, J., and Schmid, C. (2009), "Is that you? Metric learning approaches for face identification," in *2009 IEEE 12th international conference on computer vision*, IEEE, pp. 498–505.

Hadsell, R., Chopra, S., and LeCun, Y. (2006), "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, IEEE, vol. 2, pp. 1735–1742.

Hall, P. (1984), "Central limit theorem for integrated square error of multivariate nonparametric density estimators," *Journal of multivariate analysis*, 14, 1–16.

Jain, L., Mason, B., and Nowak, R. (2017), "Learning low-dimensional metrics," *arXiv preprint arXiv:1709.06171*.

Ji, Y., Sumantyo, J., Chua, M., and Waqar, M. M. (2018), "Earthquake/tsunami damage level mapping of urban areas using full polarimetric sar data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 11, 2296–2309.

Jin, R., Wang, S., and Zhou, Y. (2009), "Regularized Distance Metric Learning: Theory and Algorithm," in *NIPS*.

Ktena, S. I., Parisot, S., Ferrante, E., Rajchl, M., Lee, M., Glocker, B., and Rueckert, D. (2018), "Metric learning with spectral graph convolutions on brain connectivity networks," *NeuroImage*, 169, 431–442.

Kulis, B. (2012), "Metric learning: A survey," *Foundations and trends in machine learning*, 5, 287–364.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998), "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, 86, 2278–2324.

Lee, J. D., Lei, Q., Saunshi, N., and Zhuo, J. (2020), "Predicting what you already know helps: Provable self-supervised learning," *arXiv preprint arXiv:2008.01064*.

Li, K. (1991), "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, 86, 316–327.

Li, T., and Yuan, M. (2019), "On the optimality of Gaussian kernel based nonparametric tests against smooth alternatives," *arXiv preprint arXiv:1909.03302*.

Li, W., Zhao, R., Xiao, T., and Wang, X. (2014), "Deepreid: Deep filter pairing neural network for person re-identification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 152–159.

Liao, S., Hu, Y., Zhu, X., and Li, S. Z. (2015), "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2197–2206.

Lin, Q., Li, X., Huang, D., and Liu, J. S. (2021), "On the optimality of sliced inverse regression in high dimensions," *The Annals of Statistics*, 49, 1–20.

Ma, G., Ahmed, N. K., Willke, T. L., Sengupta, D., Cole, M. W., Turk-Browne, N. B., and Yu, P. S. (2019), "Deep graph similarity learning for brain data analysis," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2743–2751.

McArdle, B. H., and Anderson, M. J. (2001), "Fitting multivariate models to community data: a comment on distance-based redundancy analysis," *Ecology*, 82, 290–297.

Movshovitz-Attias, Y., Toshev, A., Leung, T. K., Ioffe, S., and Singh, S. (2017), "No fuss distance metric learning using proxies," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 360–368.

Musgrave, K., Belongie, S., and Lim, S. (2020), "A metric learning reality check," in *European Conference on Computer Vision*, Springer, pp. 681–699.

Oord, A., Li, Y., and Vinyals, O. (2018), "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*.

Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., and Cohen, J. P. (2020), "Revisiting training strategies and generalization performance in deep metric learning," in *International Conference on Machine Learning*, PMLR, pp. 8242–8252.

Samworth, R. J. (2012), "Optimal weighted nearest neighbour classifiers," *The Annals of Statistics*, 40, 2733–2763.

Schroff, F., Kalenichenko, D., and Philbin, J. (2015), "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013), "Equivalence of distance-based and RKHS-based statistics in hypothesis testing," *The Annals of Statistics*, 41, 2263–2291.

Sermanet, P., Lynch, C., Hsu, J., and Levine, S. (2017), "Time-contrastive networks: Self-supervised learning from multi-view observation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, pp. 486–487.

Shorten, C., and Khoshgoftaar, T. M. (2019), "A survey on image data augmentation for deep learning," *Journal of Big Data*, 6, 1–48.

Sohn, K. (2016), "Improved deep metric learning with multi-class n-pair loss objective," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1857–1865.

Székely, G. J., and Rizzo, M. L. (2005), "A new test for multivariate normality," *Journal of Multivariate Analysis*, 93, 58–80.

Tian, Y., Krishnan, D., and Isola, P. (2019), "Contrastive multiview coding," *arXiv preprint arXiv:1906.05849*.

Tian, Y., Yu, L., Chen, X., and Ganguli, S. (2020), "Understanding self-supervised learning with dual deep networks," *arXiv preprint arXiv:2010.00578*.

Tosh, C., Krishnamurthy, A., and Hsu, D. (2021), "Contrastive learning, multi-view redundancy, and linear models," in *Algorithmic Learning Theory*, PMLR, pp. 1179–1206.

Tsai, Y., Wu, Y., Salakhutdinov, R., and Morency, L. (2020), "Self-supervised learning from a multi-view perspective," *arXiv preprint arXiv:2006.05576*.

Tschannen, M., Djolonga, J., Rubenstein, P. K., Gelly, S., and Lucic, M. (2019), "On mutual information maximization for representation learning," *arXiv preprint arXiv:1907.13625*.

Wang, S., Cai, T. T., and Li, H. (2021), "Hypothesis testing for phylogenetic composition: a minimum-cost flow perspective," *Biometrika*, 108, 17–36.

Wei, C., Shen, K., Chen, Y., and Ma, T. (2020), "Theoretical analysis of self-training with deep networks on unlabeled data," *arXiv preprint arXiv:2010.03622*.

Weinberger, K. Q., and Saul, L. K. (2009), "Distance metric learning for large margin nearest neighbor classification." *Journal of machine learning research*, 10.

Xiao, H., Rasul, K., and Vollgraf, R. (2017), "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," *arXiv preprint arXiv:1708.07747*.

Xing, E. P., Ng, A. Y., Jordan, M. I., and Russell, S. (2002), "Distance metric learning with application to clustering with side-information," in *NIPS*, vol. 15, pp. 505–512.

Yao, J., Zheng, S., and Bai, Z. (2015), *Sample covariance matrices and high-dimensional data analysis*, Cambridge University Press Cambridge.

Ye, H., Zhan, D., and Jiang, Y. (2019), "Fast generalization rates for distance metric learning," *Machine Learning*, 108, 267–295.

Yi, D., Lei, Z., Liao, S., and Li, S. Z. (2014), "Deep metric learning for person re-identification," in *2014 22nd International Conference on Pattern Recognition*, IEEE, pp. 34–39.

Zhang, A., Cai, T. T., and Wu, Y. (2018), "Heteroskedastic PCA: Algorithm, optimality, and applications," *arXiv preprint arXiv:1810.08316*.

Zhang, R., Isola, P., and Efros, A. A. (2016), "Colorful image colorization," in *European conference on computer vision*, Springer, pp. 649–666.

Zhang, W., Lu, X., and Li, X. (2018), "A coarse-to-fine semi-supervised change detection for multispectral images," *IEEE Transactions on Geoscience and Remote Sensing*, 56, 3587–3599.