Neural Processing Letters (2020) 51:347–366
https://doi.org/10.1007/s11063-019-10090-0

Maximum Mean and Covariance Discrepancy for Unsupervised Domain Adaptation

Wenju Zhang[1,3] · Xiang Zhang[2,3] · Long Lan[2,3] · Zhigang Luo[1,3]

Published online: 8 August 2019

Abstract
A fundamental research topic in domain adaptation is how best to evaluate the distribution discrepancy across domains. The maximum mean discrepancy (MMD) is one of the most commonly used statistical distances in this field. However, information about distributions could be lost when adopting non-characteristic kernels by MMD. To address this issue, we devise a new distribution metric named maximum mean and covariance discrepancy (MMCD) by combining MMD and the proposed maximum covariance discrepancy (MCD). MCD probes the second-order statistics in reproducing kernel Hilbert space, which equips MMCD to capture more information compared to MMD alone. To verify the efficacy of MMCD, an unsupervised learning model based on MMCD abbreviated as McDA was proposed and efficiently optimized to resolve the domain adaptation problem. Experiments on image classification conducted on two benchmark datasets show that McDA outperforms other representative domain adaptation methods, which implies the effectiveness of MMCD in domain adaptation.

Keywords Domain adaptation · Image classification · Dimensionality reduction · Transfer learning

1 Introduction

In standard machine learning, both training and test data are assumed to be drawn from the same distribution. However, this assumption turns out unrealistic for lots of real-world

✉ Xiang Zhang
zhangxiang08@nudt.edu.cn

Long Lan
long.lan@nudt.edu.cn

Zhigang Luo
zgluo@nudt.edu.cn

1 Science and Technology on Parallel and Distributed Laboratory, National University of Defense Technology, Changsha 410073, China

2 State Key Laboratory of High Performance Computing, National University of Defense Technology, Changsha 410073, China

3 College of Computer, National University of Defense Technology, Changsha 410073, China

problems, which could cause performance degradation for testing. For example, in visual recognition task, test images may differ from training ones due to changes in backgrounds, sensors, viewpoints, etc. To address this issue, domain adaptation has been proposed in quest of adapting the models built in one domain (source domain) to serve another different but related domain (target domain), in such a way that the learned models perform well in the target domain.

In DA problems, the dataset from the source domain and the dataset from the target domain usually follow different distributions. To narrow the difference, the first thing is to appropriately evaluate the distribution discrepancy across domains for most DA models. Many candidate statistical distances, such as the Kullback–Leibler divergence [5], the Bregman divergence [33], the Wasserstein distance [32] and the maximum mean discrepancy (MMD) [11], can be used to achieve this purpose. MMD is one of the most widely used distances and it is designed to evaluate the distance between the kernel mean embedding of distributions in a reproducing kernel Hilbert space (RKHS). MMD possesses a decent theoretical property, i.e., characteristic kernels establish MMD as metrics on the space of probability distributions [9,11,34]. Due to its non-parametric form and its theoretical properties, MMD has attracted widespread attention from the DA community.

Much progress has been made in an effort to explore MMD and thereby obtain transferrable knowledge from the source domain. Theoretically, the characteristic kernel is the natural choice for MMD. However, in specific applications, non-characteristic counterparts might be more appropriate than characteristic ones [3]. Several MMD-based DA methods employ non-characteristic kernels (such as the linear kernel [16,20,25] and the polynomial kernel [1,2]) or the non-kernel linear transformation [14–16,20,37] to cope with domain shift. Despite the state-of-the-art performance achieved by the non-characteristic kernel based methods, in this paper we point out that there is still much room to improve these current methods. To be specific, with a deep insight into the recent MMD methods, we can come to the conclusion that the non-characteristic kernel based MMD could lose some statistical information that is important for DA.

To capture more information about distributions, this paper designs a new distribution metric termed the maximum mean and covariance discrepancy (MMCD). This metric is designed to address both the first- and second-order statistical information in the RKHS. Specifically, MMCD is comprised of MMD and our proposed maximum covariance discrepancy (MCD). MCD evaluates the Hilbert–Schmidt norm of the difference between covariance operators such that it addresses the second-order statistics in the RKHS. Since MMCD unites MCD and MMD, MMCD is able to simultaneously consider the first- and second-order statistics in the RKHS. Thus, when the non-characteristic kernel is used, MMCD has the ability to capture more information about distributions than MMD does. This point will be analyzed in Sects. 3.2 and 3.3.

To verify the efficacy of MMCD, we propose an unsupervised DA method that is based on MMCD (McDA) in the joint distribution adaptation (JDA) paradigm [20]. JDA has been chosen because it jointly adapts both the marginal and conditional distributions across domains. It is not easy to optimize McDA due to its non-convex term. To address this issue, we approximate the non-convex term with its convex upper bound, thereby yielding a closed-form solution to each sub-problem. Experiments in cross-domain image classification on two datasets (PIE and Office-Caltech) show the effectiveness of McDA in the JDA paradigm when compared with the baseline methods, which implies the efficacy of MMCD.

The rest of this paper is organized as follows. Section 2 reviews the related works. Section 3 introduces the proposed maximum mean and covariance discrepancy. Section 4 details the MMCD based unsupervised DA model McDA, along with an optimization algorithm to solve

it. In Sect. 5, we report the performance of McDA on two benchmark datasets. Finally, Sect. 6 concludes the paper.

2 Related Work

Domain adaptation (DA) has been widely studied in extensive literatures [26,27]. DA can be categorized into two groups supervised DA and unsupervised DA according to whether or not the data in the target domain is labeled [4,6,25,38,40]. In this paper, we focus on the unsupervised setting. To tackle the unsupervised DA problem, many methods has been proposed including sample reweighting [17], projection learning [25] and subspace alignment [8,36]. Among them, a popular way is to learn the domain-invariant representation which minimize the distribution discrepancy across domains. Notably, two methods of measuring the distribution discrepancy have been widely adopted in DA models: maximum mean discrepancy (MMD) metric [11] and the explicit order-wise distance between distributions.

As a representative metric used to measure distribution discrepancy, MMD has been widely used in DA methods. For instance, Pan et al. [25] developed transfer component analysis (TCA) to learn invariant features across domains in the reproducing kernel Hilbert space (RKHS). To investigate the benefits of conditional distributions, Long et al. [20] constructed a joint distribution adaptation (JDA) paradigm which adopted MMD in order to jointly evaluate the marginal and conditional distribution differences across domains. JDA yielded the state-of-the-art classification performance. Recently, Hsieh et al. [15] generalized JDA in order to tackle heterogeneous DA problems. Later, the closest common space learning [16] was proposed to deal with imbalanced cross-domain data problem in the frame of JDA. With advances in deep learning, MMD-based deep models have also made great success in adaptation performance across domains. Long et al. [19] proposed a deep adaptation network architecture, in which a multiple-kernel MMD was utilized in order to match cross-domain distributions of multiple task-specific CNN layers, so as to boost the transferability of deep features. Recently, Long et al. [21] have proposed a new DA method to be used in deep networks, i.e., residual transfer networks, which was able to jointly learn both adaptive classifiers and transferable features. To be different, Our MMCD explores both the first- and second-order statistics in the RKHS.

In addition to MMD-based models, recent work has explicitly matched the order-wise statistics of cross-domain distributions, thereby taking the statistics of distributions into account. For example, Sun et al. [35] proposed an efficient unsupervised DA method, which they designated as correlation alignment, to align the covariance matrix of the source and target distributions. Jiang et al. [18] proposed to use the second moment matching so as to learn the domain-invariant feature. Zong et al. [41] took a different approach which explores both the least-square regression and mean–covariance feature matching in order to enhance the performance of cross-corpus speech emotion recognition. Zellinger et al. [39] took a step forward to explore higher order moment which learns domain-invariant representations by means of explicitly order-wise moment matching. Different from the above mentioned methods which all study in the original feature space, our MMCD, however, explores the higher order statistics in the RKHS by using the kernel trick.

It is noted that MMCD is related to the work of [22] which embeds distributions in a finite dimensional feature space, and matches the mean and covariance feature statistics in order to train generative adversarial networks. The proposed MMCD differs from this approach in two ways. Firstly, MMCD function set is from the RKHS which is infinitely dimensional in

the setting of PSD kernels. In comparison, the function set used by [22] is parameterized by a finite-dimensional CNN. Secondly, MCD can be represented as the Hilbert–Schmidt norm. By contrast, the covariance matching in [22] is related to the nuclear norm. Besides, we notice that the empirical estimator of squared MCD has similar form to that from [28,29]. However, they are different in three aspects. Firstly, MCD is defined in the frame of integral probability metrics [24] and can be regarded as the second-order generation of MMD, while [28] uses the empirical estimator of squared MCD directly. Secondly, MCD serves as distribution metric, while [28] uses covariance operators as image representation. Thirdly, this paper unites both MMD and MCD as MMCD to measure the distribution discrepancy.

3 Maximum Mean and Covariance Discrepancy

This section will introduce the maximum covariance discrepancy (MCD) and its joint variant the maximum mean and covariance discrepancy (MMCD). In addition, several theoretical analyses will be provided. At first, we introduce the notations that are used throughout this section.

Let $\mathcal{H}$ be a reproducing kernel Hilbert space (RKHS) over X with associated kernel $k(\cdot,\cdot)$, whose canonical feature map is $\phi(x) = k(x,\cdot)$ for $x \in X$. Let x and y be random variables defined on X, we assume $x \sim p$ and $y \sim q$, where $x \sim p$ indicates x follows distribution p. The notation $E_{x\sim p}$ denotes the expectation with respect to p, and we abbreviate $E_{x\sim p}$ and $E_{x\sim q}$ to E_x and E_y respectively when there is no ambiguity. Following [12], when given $f, g \in \mathcal{H}$, the tensor product operator $f \otimes g : \mathcal{H} \to \mathcal{H}$ is defined as $(f \otimes g)\, h = f\langle g, h\rangle_{\mathcal{H}}$ for all $h \in H$, where $\otimes$ denotes the tensor product. Based on the tensor product operator and probability distributions, the centered covariance operator $C : \mathcal{H} \to \mathcal{H}$ associated with the distribution p has the definition $C[p] = E_{x\sim p}[\phi(x) \otimes \phi(x)] - E_{x\sim p}[\phi(x)] \otimes E_{x\sim p}[\phi(x)]$. At last, we make two assumptions in our paper. First, we assume the kernel $k(\cdot,\cdot)$ to be bounded, which makes the covariance operator bounded and the mean embedding of distributions in $\mathcal{H}$ exists [11]. Second, we assume $\mathcal{H}$ to be separable such that the basis of $\mathcal{H}$ is finite or countably infinite.

3.1 Definition

With the assumptions and notations above, we define the MCD in Definition 1.

Definition 1 The maximum covariance discrepancy (MCD) is defined as

$$\mathrm{MCD}[p, q, \mathcal{H}] = \sup_{\|a\|\le 1} \sum_{i,j\in I} a_{ij} \left(\mathrm{cov}\left[e_i(x), e_j(x)\right] - \mathrm{cov}\left[e_i(y), e_j(y)\right]\right), \tag{1}$$

where $\{e_i | i \in I\}$ is an orthogonal basis of $\mathcal{H}$, $\|a\| = (\sum_{i,j\in I} a_{ij}^2)^{1/2}$, and the notation cov has the following formula: $\mathrm{cov}\left[e_i(x), e_j(x)\right] = E_x\left[e_i(x)\, e_j(x)\right] - E_x[e_i(x)]\, E_x\left[e_j(x)\right]$.

Next, we show that the (1) can be associated with the Hilbert–Schmidt norm with the following lemma:

Lemma 1

$$\mathrm{MCD}[p, q, \mathcal{H}] = \|C[p] - C[q]\|_{\mathrm{HS}}, \tag{2}$$

where $\|\|_{\mathrm{HS}}$ *denotes the Hilbert–Schmidt norm of the vectors in* HS $(\mathcal{H})$*, which is the Hilbert space of Hilbert–Schmidt operators mapping from* $\mathcal{H}$ *to* $\mathcal{H}$*.*

Proof By using the reproducing property of the kernel k, we have

$$\begin{aligned}&\sum_{i,j\in I} a_{ij}\left(E_x\left[e_i(x)\,e_j(x)\right]-E_x\left[e_i(x)\right]E_x\left[e_j(x)\right]\right)\\&=\left\langle\sum\nolimits_{i,j\in I} a_{ij}e_i\otimes e_j, C[p]\right\rangle_{\mathrm{HS}}.\end{aligned}\tag{3}$$

Similarly, we also have

$$\begin{aligned}&\sum_{i,j\in I} a_{ij}\left(E_y\left[e_i(y)\,e_j(y)\right]-E_y\left[e_i(y)\right]E_y\left[e_j(y)\right]\right)\\&=\left\langle\sum\nolimits_{i,j\in I} a_{ij}e_i\otimes e_j, C[q]\right\rangle_{\mathrm{HS}}.\end{aligned}\tag{4}$$

In terms of (3) and (4), we have

$$\begin{aligned}\mathrm{MCD}[p,q,\mathcal{H}]&=\sup_{\|a\|\le 1}\left\langle\sum\nolimits_{i,j\in I} a_{ij}e_i\otimes e_j, C[p]-C[q]\right\rangle_{\mathrm{HS}}\\&=\sup_{h\in\mathcal{H}\otimes\mathcal{H},\ \|h\|_{\mathcal{H}\otimes\mathcal{H}}\le 1}\left\langle h^*, C[p]-C[q]\right\rangle_{\mathrm{HS}}\\&=\sup_{h\in\mathcal{H}\otimes\mathcal{H},\ \|h^*\|_{\mathrm{HS}}\le 1}\left\langle h^*, C[p]-C[q]\right\rangle_{\mathrm{HS}}\\&=\|C[p]-C[q]\|_{\mathrm{HS}},\end{aligned}\tag{5}$$

where h^* denotes the tensor product operator of h which belongs to the tensor product space $\mathcal{H}\otimes\mathcal{H}$. Since $\{e_i|i\in I\}$ is the orthogonal basis of $\mathcal{H}$, $\left\{e_i\otimes e_j|i,j\in I\right\}$ consists of the orthogonal basis of $\mathcal{H}\otimes\mathcal{H}$. Thus, the second equality holds. The third equality follows from the following fact:

$$\begin{aligned}\left\|h^*\right\|_{\mathrm{HS}}^2&=\left\langle\sum\nolimits_{i,j\in I} a_{ij}e_i\otimes e_j, \sum\nolimits_{i,j\in I} a_{ij}e_i\otimes e_j\right\rangle_{\mathrm{HS}}\\&=\sum_{i,j\in I} a_{ij}\sum_{m,n\in I} a_{mn}\left\langle e_i\otimes e_j, e_m\otimes e_n\right\rangle_{\mathrm{HS}}\\&=\sum_{i,j\in I} a_{ij}\sum_{m,n\in I} a_{mn}\langle e_i,e_m\rangle_{\mathcal{H}}\left\langle e_j,e_n\right\rangle_{\mathcal{H}}\\&=\sum_{i,j\in I} a_{ij}^2\\&=\|h\|_{\mathcal{H}\otimes\mathcal{H}}^2.\end{aligned}\tag{6}$$

□

To compute it, we express MCD as the expansion formula with respect to kernel functions, using the following lemma:

Lemma 2 *The squared* MCD *terms of kernel functions is*

$$\begin{aligned}\mathrm{MCD}^2[p,q,\mathcal{H}]&=E_{x,x'}\left[k^2\left(x,x'\right)\right]-2E_x\left[E_{x'}^2\left[k\left(x,x'\right)\right]\right]\\&+E_{x,x'}^2\left[k\left(x,x'\right)\right]-2E_{x,y}\left[k^2(x,y)\right]+2E_x\left[E_y^2\left[k(x,y)\right]\right]\\&+2E_y\left[E_x^2\left[k(x,y)\right]\right]-2E_{x,y}^2\left[k(x,y)\right]+E_{y,y'}\left[k^2\left(y,y'\right)\right]\\&-2E_y\left[E_{y'}^2\left[k\left(y,y'\right)\right]\right]+E_{y,y'}^2\left[k\left(y,y'\right)\right],\end{aligned}\tag{7}$$

where x' and y' are independent copy of x and y, with the same distribution, respectively.

Given the limited observations $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ sampled from p and q, and based on Lemma 2 and statistical theory, an empirical estimator of squared MCD can be given by

$$\widehat{\mathrm{MCD}}^2[X, Y, \mathcal{H}] = \frac{1}{n^2} tr\left(K_{XX} L_n K_{XX} L_n\right) - \frac{2}{nm} tr\left(K_{XY} L_m K_{XY}^T L_n\right) + \frac{1}{m^2} tr\left(K_{YY} L_m K_{YY} L_m\right), \tag{8}$$

where $(K_{XX})_{ij} = k(x_i, x_j)$, $(K_{XY})_{ij} = k(x_i, y_j)$, $(K_{YY})_{ij} = k(y_i, y_j)$, and $L_n = I_n - \frac{1}{n} 1_n^T 1_n$ wherein I_n is the identity matrix of size n, and 1_n is the vector of ones with length n. This can be notated as, $L_m = I_m - \frac{1}{m} 1_m^T 1_m$. To be concise, we can rewrite (8) as the elegant formula

$$\widehat{\mathrm{MCD}}^2[X, Y, \mathcal{H}] = tr(KZKZ), \tag{9}$$

where

$$K = \begin{bmatrix} K_{XX} & K_{XY} \\ K_{YX} & K_{YY} \end{bmatrix}, \tag{10}$$

and

$$Z_{ij} = \begin{cases} \frac{1}{n} - \frac{1}{n^2}, & i = j, x_i \in X \\ -\frac{1}{n^2}, & i \neq j, x_i \in X, x_j \in X \\ \frac{1}{m^2} - \frac{1}{m}, & i = j, x_i \in Y \\ \frac{1}{m^2}, & i \neq j, x_i \in Y, x_j \in Y \\ 0, & otherwise. \end{cases} \tag{11}$$

Moreover, we can also yield another equivalent representation of (8), as follows:

$$\left\| \frac{1}{n} \phi(X) L_n \phi(X)^T - \frac{1}{m} \phi(Y) L_m \phi(Y)^T \right\|_{\mathrm{HS}}^2, \tag{12}$$

where $k(x, y) = \phi(x)^T \phi(y)$, wherein $\phi(X) = [\phi(x_1), \ldots, \phi(x_n)]$ and $\phi(Y) = [\phi(y_1), \ldots, \phi(y_m)]$. This form will be utilized in the later section.

Now, we unite MMD and MCD into a joint metric, namely maximum mean and covariance discrepancy (MMCD) in order to capture more statistical information from data distributions.

Definition 2 The maximum mean and covariance discrepancy (MMCD) is defined as

$$\mathrm{MMCD}[p, q, \mathcal{H}] = \left(\|\mu[p] - \mu[q]\|_{\mathcal{H}}^2 + \beta \|C[p] - C[q]\|_{\mathrm{HS}}^2 \right)^{1/2}, \tag{13}$$

where $\mu[p] = E_x[\phi(x)]$ and β is a non-negative parameter.

According to this definition, MMCD can be proven as a distribution metric when the associate kernel of $\mathcal{H}$ is characteristic. This property can be established by Theorem 1.

Theorem 1 *Let the associated kernel of $\mathcal{H}$ be characteristic. Then* $\mathrm{MMCD}[p, q, \mathcal{H}] = 0$ *if and only if $p = q$. Moreover,* $\mathrm{MMCD}[p, q, \mathcal{H}]$ *is a metric on the space of probability distribution.*

Proof According to [7], a metric d should satisfy four properties: non-negativity, $d(p,q) = 0 \Leftrightarrow p = q$, symmetry and triangle inequality. Apparently, MMCD is non-negative due to the non-negativity of the norm. For the necessity and sufficiency of the second property, first, under the condition $p = q$, $\mathrm{MMCD}[p,q,\mathcal{H}] = 0$; conversely, $p = q$ can be deduced according to the definition of characteristic kernel [9] and $\mathrm{MMD}[p,q,\mathcal{H}] = 0$. As a side assertion, it is easy to verify that $\mathrm{MMCD}[p,q,\mathcal{H}] = \mathrm{MMCD}[q,p,\mathcal{H}]$ which demonstrates the symmetric property of MMCD. The only property left to prove is the triangle inequality. First, we need to prove that the MCD meets this property. Let $\mathrm{d}_{ij}(x,y) = \mathrm{cov}[e_i(x), e_j(x)] - \mathrm{cov}[e_i(y), e_j(y)]$, and then we have

$$\begin{aligned}
&\mathrm{MCD}[p,r,\mathcal{H}] + \mathrm{MCD}[r,q,\mathcal{H}] \\
&= \sup_{\|a\|\le 1}\left(\sum_{i,j\in I} a_{ij}\mathrm{d}_{ij}(x,z)\right) + \sup_{\|a\|\le 1}\left(\sum_{i,j\in I} a_{ij}\mathrm{d}_{ij}(z,y)\right) \\
&\ge \sup_{\|a\|\le 1}\left(\sum_{i,j\in I} a_{ij}\mathrm{d}_{ij}(x,z) + \sum_{i,j\in I} a_{ij}\mathrm{d}_{ij}(z,y)\right) \\
&= \sup_{\|a\|\le 1}\left(\sum_{i,j\in I} a_{ij}\mathrm{d}_{ij}(x,y)\right) \\
&= \mathrm{MCD}[p,q,\mathcal{H}],
\end{aligned} \tag{14}$$

where $z \sim r$. Next, we prove the triangle inequality holds for the MMCD. To simplify the notation, let $\mathrm{M}[p,q] = \mathrm{MMD}[p,q,\mathcal{H}]$, $\mathrm{C}[p,q] = \mathrm{MCD}[p,q,\mathcal{H}]$ and $\mathrm{MC}[p,q] = \mathrm{MMCD}[p,q,\mathcal{H}]$, and then we have

$$\begin{aligned}
\mathrm{MC}^2[p,q] &= \mathrm{M}^2[p,q] + \beta\mathrm{C}^2[p,q] \\
&\le (\mathrm{M}[p,r] + \mathrm{M}[r,q])^2 + \beta(\mathrm{C}[p,r] + \mathrm{C}[r,q])^2 \\
&= \mathrm{M}^2[p,r] + \beta\mathrm{C}^2[p,r] + \mathrm{M}^2[r,q] + \beta\mathrm{C}^2[r,q] \\
&\quad + 2\left(\mathrm{M}[p,r]\,\mathrm{M}[r,q] + \sqrt{\beta}\mathrm{C}[p,r]\sqrt{\beta}\mathrm{C}[r,q]\right) \\
&\le \mathrm{M}^2[p,r] + \beta\mathrm{C}^2[p,r] + \mathrm{M}^2[r,q] + \beta\mathrm{C}^2[r,q] \\
&\quad + 2\left(\mathrm{M}^2[p,r] + \beta\mathrm{C}[p,r]\right)^{1/2}\left(\mathrm{M}^2[r,q] + \beta\mathrm{C}[r,q]\right)^{1/2} \\
&= (\mathrm{MC}[p,r] + \mathrm{MC}[r,q])^2,
\end{aligned} \tag{15}$$

where the first inequality holds because both MMD and MCD meet the triangle inequality, and the second inequality holds from the Cauchy–Schwarz inequality. Taking the square root of both sides, there holds $\mathrm{MMCD}[p,q] \le \mathrm{MMCD}[p,r] + \mathrm{MMCD}[r,q]$. Obviously, MMCD meets the metric definition, and is, therefore, a metric. □

In fact, while Theorem 1 relies on the condition that the kernel is characteristic, even if the condition does not hold, MMCD is still a pseudo-metric. According to (9) and [11], the empirical estimator of the squared MMCD can be given by

$$\widehat{\mathrm{MMCD}}^2[p,q,\mathcal{H}] = tr(KM) + \beta tr(KZKZ), \tag{16}$$

and

$$\widehat{\mathrm{MMCD}}^2[p,q,\mathcal{H}] = \left\| \frac{1}{n}\phi(X)\,1_n - \frac{1}{m}\phi(Y)\,1_m \right\|_{\mathcal{H}}^2 + \beta \left\| \frac{1}{n}\phi(X)\,H_n\phi(X)^T - \frac{1}{m}\phi(Y)\,H_m\phi(Y)^T \right\|_{\mathrm{HS}}^2, \tag{17}$$

where

$$M_{ij} = \begin{cases} \frac{1}{n^2}, & x_i, x_j \in X \\ \frac{1}{m^2}, & x_i, x_j \in Y \\ -\frac{1}{nm}, & otherwise. \end{cases} \tag{18}$$

3.2 Explicit Representation of MMCD

As stated above, we have defined MMCD as the (pseudo-) metric of distributions. However, it is not easy to understand what specific information about the distribution is captured by MMCD. Towards this goal, we deduce the explicit representation of MMCD with the polynomial kernel and the linear kernel to illustrate the mechanism of MMCD.

We first introduce the explicit representation of MMCD associated with the polynomial kernel with a specific degree d, i.e., $k(x, y) = (x^T y + c)^d$. According to the explicit feature map of the polynomial kernel [23,31], MMCD can be explicitly written as

$$\mathrm{MMCD}^2[p,q] = \left\| E\left[W_p\right] - E\left[W_q\right] \right\|_2^2 + \beta \left\| E\left[W_p W_p^T\right] - E\left[W_p\right] E\left[W_p\right]^T - \left(E\left[W_q W_q^T\right] - E\left[W_q\right] E\left[W_q\right]^T \right) \right\|_F^2, \tag{19}$$

where

$$W_p = \left[\sqrt{c^{d-1} C_d^1}\, vec(\otimes^1 x)^T, \ldots, \sqrt{c^{d-i} C_d^i}\, vec(\otimes^i x)^T, \ldots, \sqrt{c^0 C_d^d}\, vec(\otimes^d x)^T \right]^T, \tag{20}$$

and C_d^i denotes the binomial coefficient, $vec()$ converts the matrix into a column vector, and $\otimes^i x$ denotes the ith order tensor product of x, wherein $x \sim p$.

By setting $d = 1$, (19) becomes

$$\mathrm{MMCD}^2[p,q] = \|E[x] - E[y]\|_2^2 + \beta \left\| E\left[xx^T\right] - E[x]\,E[x]^T - \left(E\left[yy^T\right] - E[y]\,E[y]^T \right) \right\|_F^2, \tag{21}$$

where $x \sim p$ and $y \sim q$. Given limited X and Y sampled from p and q, respectively, there is

$$\widehat{\mathrm{MMCD}}^2[p,q] = \left\| \mu_p - \mu_q \right\|_2^2 + \beta \left\| \Sigma_p - \Sigma_q \right\|_F^2, \tag{22}$$

where $\mu_p = \frac{1}{n} X 1_n$ and $\Sigma_p = \frac{1}{n} X H_n X^T$ are mean vector and covariance matrix of X, respectively. Specially, when $c = 0$, the polynomial kernel becomes the linear kernel $k(x, y) = x^T y$. Accordingly, (21) is exactly the explicit representation of MMCD with the linear kernel, and (22) is still the corresponding empirical estimator. As a result, both (21) and (22) show that MMD and MCD of MMCD measures the difference between means

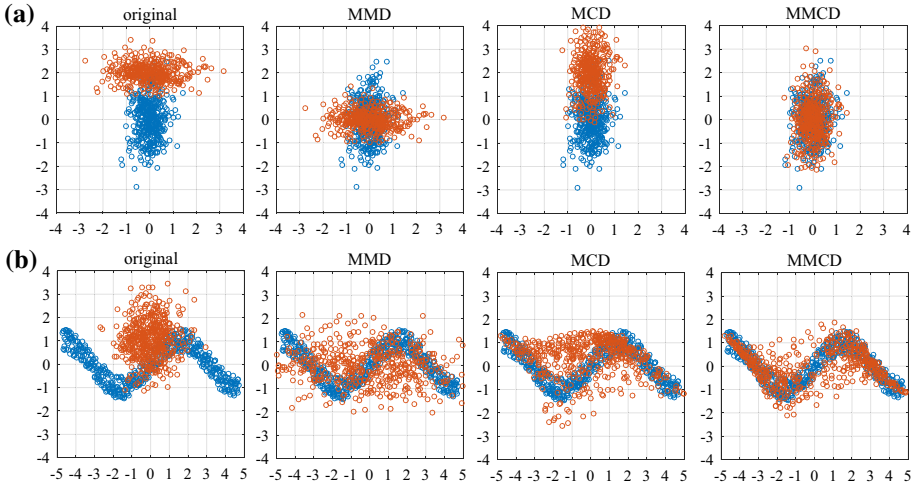

Fig. 1 Two toy examples for comparison of MMD, MCD and MMCD on the synthetic data. **a** Blue and red points are sampled from two Gaussian distributions whose both the mean and covariance are different, and the linear kernel is used, **b** blue points are sampled from a complicated distribution with the zero mean (0, 0), while red points are sampled from a Gaussian distribution with the mean (0, 1). The polynomial kernel with degree two is used therein. (Color figure online)

and covariances of distributions, respectively, when the polynomial kernel with degree $d = 1$ or the linear kernel is adopted.

Moreover, by setting $d = 2$, there is

$$W_p = \left[\sqrt{2c}vec\left(xx^T\right)^T, x^T\right]^T. \tag{23}$$

Combination of (19) with (23) shows that the MMD term of MMCD measures the difference between both the first and second raw moments of two distributions, whereas the MCD term measures the difference between covariances of up to the second raw moment of two distributions. An analogy for setting $d > 2$ means that the MCD term can capture higher order statistical information. From the statistical perspective, more insights into MCD remain to be further investigated in the future work.

3.3 Toy Examples

To clarify the efficacy of MMCD, which fuses both first- and second-order statistics in the RKHS, we illustrate two toy examples in Fig. 1a, b. We synthesize two groups of data points which follow different distributions in the two-dimensional space. Then, we fix one group of data points (in blue) and adjust the points from the other group (in red) by minimizing the values of MMD, MCD and MMCD via the gradient descent algorithm. Two types of kernels, that is, the linear kernel (Fig. 1a) and the polynomial kernel with $d = 2$ and $c = 1$ (Fig. 1b), are adopted. For concise, the details of computing their gradients are left in "Appendix A".

From Fig. 1a, through the distribution matching of MMD, the means of the red and blue points are almost identical, but the corresponding spread of these data points has different shapes. That is, MMD cannot match the covariances of two distributions. But, for MCD, the result is reversed, namely, the covariances of two groups of data points appear to be similar,

but the corresponding means of these data points keep invariant as before and still differ from each other. Notably, after MMCD performs the distribution matching over two groups of data points, the distribution regions of both the red and blue points overlap highly. This is because MMCD simultaneously takes both mean and covariance differences into consideration, while either MMD or MCD could singly consider one aspect where the linear kernel is used.

Figure 1b displays, for the complicated distribution, the distribution matching results of MMD, MCD and MMCD, when the polynomial kernel with degree two is equipped. In this case, MMD induces different spread shapes, which corresponds to the high-order statistics. Meanwhile, it can be verified that the mean of the red data points is above that of the blue points which is zero. Thus, both MMD and MCD fail to make two distributions matched in this example. By using MMCD, the red and blue points become highly overlapped. This implies the efficacy of MMCD which explores both first- and second-order statistics in the RKHS, as compared to MMD and MCD. Hence, in contrast with MMD and MCD, MMCD has the promising potential of matching data distributions across domains for domain adaptation. More empirical analyses about the efficacy of MMCD in domain adaptation are shown in experiments.

4 Domain Adaptation Via MMCD

In this section, we apply the MMCD to unsupervised domain adaptation problem to verify the efficacy of MMCD. Unsupervised domain adaptation is still very challenging, as there is no supervised knowledge of target domains. Recently, joint distribution adaptation (JDA) [20] has been proven a promising domain adaptation method which matches both marginal and conditional distributions. JDA adopts MMD as the distance measurement of distributions; however, in the case of non-characteristic kernels, MMD based JDA loses the high-order statistical information. As mentioned in the previous section, MMCD has a better chance to capture more information about distributions than MMD. We thus substitute our MMCD for MMD in the frame of JDA to produce a new unsupervised domain adaptation method called McDA. In the subsection, we first introduce the problem formulation and notation, and then propose the McDA.

4.1 Problem Formulation and Notations

Given the dataset $D_s = \{x_1, \ldots, x_{n_s}\}$ in the source domain with labels $y_1, \ldots, y_{n_s}$ and the unlabeled dataset $D_t = \{x_{n_s+1}, \ldots, x_{n_s+n_t}\}$ in the target domain under the assumption that both the marginal and conditional probability distributions in two domains are different, i.e., $P_s(x_s) \neq P_t(x_t)$ and $P_s(y_s|x_s) \neq P_t(y_t|x_t)$. We will adapt both marginal and conditional distributions in order to train a robust classifier on target dataset by leveraging the labeled source dataset.

For clarity, we now summarize several commonly used notations. The source and target data are denoted as $X_s = [x_1, \ldots, x_{n_s}]$ and $X_t = [x_{n_s+1}, \ldots, x_{n_s+n_t}]$, respectively, and, for brevity, $X = [X_s, X_t]$. The centeralized matrix L_n is defined as $L_n = I_n - \frac{1}{n} 1_n^T 1_n$, where I_n is the identity matrix of size n, and 1_n is the vector of ones with length n.

Since our model McDA is based on the JDA paradigm, it is necessary to review it before our text. JDA is to adapt both marginal and conditional distributions across domains by minimizing the following objective:

$$\min_{\phi} \|E[\phi(x_s)] - E[\phi(x_t)]\|^2 + \|E[y_s|\phi(x_s)] - E[y_s|\phi(x_s)]\|^2. \tag{24}$$

According to [20], ϕ can serve as a linear transformation, or a non-linear feature map associated with the kernel. The property of MMD still is an open problem when substituting its feature map for linear transformation. Then we roughly regard the terms of (24) as the estimation of the generalized MMD whose feature map can be replaced with any arbitrary maps. Thus, following [20], we obtain two variants of our McDA according to the used feature maps: the linear transformation and the kernel feature map.

4.2 McDA

For the first variant of McDA, we adopt the linear transformation denoted by a matrix $A \in R^{m\times k}$ to MMCD. According to (17), the distance between marginal distributions across domains can be converted into

$$\begin{aligned}&\left\|\frac{1}{n_s}A^T X_s 1_{n_s} - \frac{1}{n_t}A^T X_t 1_{n_t}\right\|_2^2 + \beta\left\|\frac{1}{n_s}A^T X_s L_{n_s} X_s^T A - \frac{1}{n_t}A^T X_t L_{n_t} X_t^T A\right\|_F^2 \\ &= tr\left(A^T X M_0 X^T A\right) + \beta\left\|A^T X Z_0 X^T A\right\|_F^2,\end{aligned} \tag{25}$$

where M_0 is defined as:

$$(M_0)_{ij} = \begin{cases}\frac{1}{n_s^2}, & x_i, x_j \in D_s \\ \frac{1}{n_t^2}, & x_i, x_j \in D_t \\ -\frac{1}{n_s n_t}, & otherwise,\end{cases} \tag{26}$$

and Z_0, termed as the MCD matrix, is defined as:

$$(Z_0)_{ij} = \begin{cases}\frac{1}{n_s} - \frac{1}{n_s^2}, & i = j, x_i \in D_s \\ -\frac{1}{n_s^2}, & i \neq j, x_i \in D_s, x_j \in D_s \\ \frac{1}{n_t^2} - \frac{1}{n_t}, & i = j, x_i \in D_t \\ \frac{1}{n_t^2}, & i \neq j, x_i \in D_t, x_j \in D_t \\ 0, & otherwise.\end{cases} \tag{27}$$

Similarly, the discrepancy between conditional distributions across transformed domains can be cast as:

$$tr\left(A^T X M_c X^T A\right) + \beta\left\|A^T X Z_c X^T A\right\|_F^2, \tag{28}$$

where $c \in \{1, \ldots, C\}$, and M_c is

$$(M_c)_{ij} = \begin{cases}\frac{1}{n_{s,c}^2}, & x_i \in D_{s,c}, x_j \in D_{s,c} \\ \frac{1}{n_{t,c}^2}, & x_i \in D_{t,c}, x_j \in D_{t,c} \\ -\frac{1}{n_{s,c} n_{t,c}}, & \begin{cases}x_i \in D_{s,c}, x_j \in D_{t,c} \\ x_j \in D_{s,c}, x_i \in D_{t,c}\end{cases} \\ 0, & otherwise,\end{cases} \tag{29}$$

wherein $D_{s,c}\left(D_{t,c}\right) = \{x_i | x_i \in D_s\left(D_t\right), y_i = c\}$ and $n_{s,c}\left(n_{t,c}\right) = \left|D_{s,c}\left(D_{t,c}\right)\right|$, while Z_c is defined as

$$(Z_c)_{ij} = \begin{cases} \frac{1}{n_{s,c}} - \frac{1}{n_{s,c}^2}, & i = j, x_i \in D_{s,c} \\ -\frac{1}{n_{s,c}^2}, & i \neq j, x_i \in D_{s,c}, x_j \in D_{s,c} \\ \frac{1}{n_{t,c}^2} - \frac{1}{n_{t,c}}, & i = j, x_i \in D_{t,c} \\ \frac{1}{n_{t,c}^2}, & i \neq j, x_i \in D_{t,c}, x_j \in D_{t,c} \\ 0, & otherwise. \end{cases} \tag{30}$$

By combining (25) and (28) and denoting $H = L_{n_s+n_t}$, we can obtain the overall objective function as follows:

$$\min_{A^T X H X^T A = I} \sum_{c=0}^{C} tr\left(A^T X M_c X^T A\right) + \beta \sum_{c=0}^{C} \left\| A^T X Z_c X^T A \right\|_F^2 + \lambda \|A\|_F^2, \tag{31}$$

where I is the identity matrix of size k, a constraint is imposed to avoid yielding a trivial solution and the parameter λ remains (31) as a well-defined optimization problem.

For the second variant of McDA, we, however, adopt the kernel feature map $\varphi: x \rightarrow \varphi(x)$ to MMCD. Let the kernel matrix $K = \varphi(X)^T \varphi(X)$. The MMCD-based distance of both marginal and conditional cross-domain distributions is defined as:

$$\sum_{c=0}^{C} tr\left(K M_c\right) + \beta \sum_{c=0}^{C} tr\left(K Z_c K Z_c\right). \tag{32}$$

Following [25], we employ the empirical kernel mapping $K = (K K^{-1/2})(K^{-1/2} K)$ and further introduce the low dimensional transformation in order to obtain $\tilde{K} = (K K^{-1/2}\tilde{A})(\tilde{A}^T K^{-1/2} K) = K A A^T K$, where $A = K^{-1/2}\tilde{A}$. Thus, substituting $\tilde{K}$ into (32) leads to the following minimization problem

$$\min_{A^T K H K^T A = I} \sum_{c=0}^{C} tr\left(A^T K M_c K^T A\right) + \beta \sum_{c=0}^{C} \left\| A^T K Z_c K^T A \right\|_F^2 + \lambda \|A\|_F^2. \tag{33}$$

4.3 Optimization Algorithm

Both (31) and (33) possess similar optimization problems; hence, only the optimization algorithm for (31) is provided here. Obviously, (31) is non-convex with the variable A and hard to optimize, as it contains a non-convex fourth-order term which is the second term of (31). In [20,25], the optimization problem have the closed-form solution by solving a generalized eigen-decomposition problem. To preserve this property, we can approximate the fourth-order term in (31) with its convex upper bound by using the following theorem:

Theorem 2 *The following inequality holds*

$$\sum_{c=0}^{C} \left\| A^T X Z_c X^T A \right\|_F^2 \leq \sigma k \sum_{c=0}^{C} \left\| A^T X Z_c X^T \right\|_F^2, \tag{34}$$

where k is the reduced dimensionality and $\sigma = \left\| (X H X^T)^{-1/2} \right\|^2$.

Proof

$$\begin{aligned}\sum_{c=0}^{C}\left\|A^T X Z_c X^T A\right\|_F^2 &= \sum_{c=0}^{C}\left\|A^T X Z_c X^T \left(XHX^T\right)^{-1/2}\left(XHX^T\right)^{1/2} A\right\|_F^2 \\ &\leq \sum_{c=0}^{C}\left\|A^T X Z_c X^T \left(XHX^T\right)^{-1/2}\right\|_F^2 \left\|\left(XHX^T\right)^{1/2} A\right\|_F^2 \\ &= k\sum_{c=0}^{C}\left\|A^T X Z_c X^T \left(XHX^T\right)^{-1/2}\right\|_F^2 \\ &\leq k\sum_{c=0}^{C}\left\|A^T X Z_c X^T\right\|_F^2 \left\|\left(XHX^T\right)^{-1/2}\right\|_F^2 \\ &= k\sigma\sum_{c=0}^{C}\left\|A^T X Z_c X^T\right\|_F^2. \end{aligned} \tag{35}$$

The first equation holds because XHX^T is semi-definite positive, while the second and forth inequalities follow the Cauchy–Schwarz inequality. In terms of the constraint of (31), $k = Tr\left(A^T XHX^T A\right) = Tr\left(I_k\right)$. □

According to Theorem 2, we absorb the constant of (34) in order to obtain the convex objective of (31), as follows:

$$\min_{A^T XHX^T A=I}\sum_{c=0}^{C} tr\left(A^T X M_c X^T A\right) + \beta\sum_{c=0}^{C}\left\|A^T X Z_c X^T\right\|_F^2 + \lambda\left\|A\right\|_F^2. \tag{36}$$

Then, we derive the Lagrange function of (36), as follows:

$$\begin{aligned} &tr\left(A^T\left(X\sum_{c=0}^{C} M_c X^T + \beta\sum_{c=0}^{C} X Z_c X^T X Z_c X^T + \lambda I\right)A\right) \\ &+ tr((I - A^T XHX^T A)\Phi), \end{aligned} \tag{37}$$

where the diagonal matrix Φ denotes the Lagrange multipliers. By setting the derivative of (37) over A to zero, we can obtain the following generalized eigenvalue decomposition problem:

$$\left(X\sum_{c=0}^{C} M_c X^T + \beta\sum_{c=0}^{C} X Z_c X^T X Z_c X^T + \lambda I\right)A = XHX^T\Phi. \tag{38}$$

Following the JDA paradigm [20], we learn the transformation matrix via (38) and can then project all the samples to a low-dimensional subspace. Based on the labeled source data, we can train a specific classifier to assign pseudo-labels to the samples of the target domain, and then repeat the procedure above until convergence. The overall procedure of McDA is summarized in Algorithm 1.

Algorithm 1: McDA

Input: X, y_s, k, λ, β, and N
Output: A
Construct M_0 by (26) and Z_0 by (27), set $M_c = 0$ and $Z_c = 0$ for $c = 1, ..., C$;
for $i = 1 : N$ **do**
  Solve (38) and select the k smallest eigenvectors to construct A;
  Update pseudo-target labels using trained classifier;
  Construct M_c and Z_c by (29) and (30) respectively for $c = 1, ..., C$.
end

5 Experiments

This section is to verify the effectiveness of McDA by comparing its classification performance against the baseline methods on two benchmark datasets including PIE and Office-Caltech.

5.1 Datasets

The PIE dataset contains face images of size 32x32 from 68 individuals with different poses, illuminations and expressions. Following [20], We evaluate our method using five subsets from the PIE face dataset. Each subset corresponds to a different pose, e.g. P1 (left), P2 (upward), P3 (downward), P4 (frontal) and P5 (right). We then construct 20 cross-domain datasets via a pairwise combination of subsets, i.e., P1 $\rightarrow$ P2, P1 $\rightarrow$ P3, . . ., P5 $\rightarrow$ P4. Since the source and target face images from each cross-domain dataset have different poses, they will follow different distributions.

Office-Caltech is a widely used dataset for domain adaptation which contains four domains: A (Amazon), W (Webcam), D (DSLR) and C (Caltech-256) [13,30]. The SURF feature is extracted for each image before being converted into histograms over an 800-bin codebook clustered by k-means on the Amazon database. We construct 12 cross-domain datasets by pairwise combination of the four domains, i.e., C $\rightarrow$ A, C $\rightarrow$ W, . . ., D $\rightarrow$ W. Several sample images from PIE and Office-Caltech datasets are illustrated in Fig. 2.

5.2 Cross-Domain Image Classification

We compare McDA with five typical and related baseline methods including nearest neighbor classifier (NN), principal component analysis (PCA), correlation alignment (CA) [35], transfer component analysis (TCA) [25], geodesic flow kernel (GFK) [10], and joint domain adaptation (JDA) [20]. As suggested by [10,20], NN serves as the base classifier for both McDA and compared methods.

We follow the evaluation protocol of [20] for both benchmark datasets. In order to give a fair comparison, we also give the candidate ranges for the parameters used in the compared methods. Both TCA and JDA have a regularization parameter, and its candidate values are as follows: {0.001, 0.01, 0.1, 1, 10, 100}. For McDA, it is rather expensive to search all the possible parameter combinations for the best average accuracy. Since McDA shares the reduced dimensionality k and the regularization parameter λ with JDA, we directly set λ in McDA to the specific value at which JDA achieves its best accuracy and then search the other parameter β from {0.001, 0.01, 0.1, 1, 10, 100} at each k. We further empirically set the

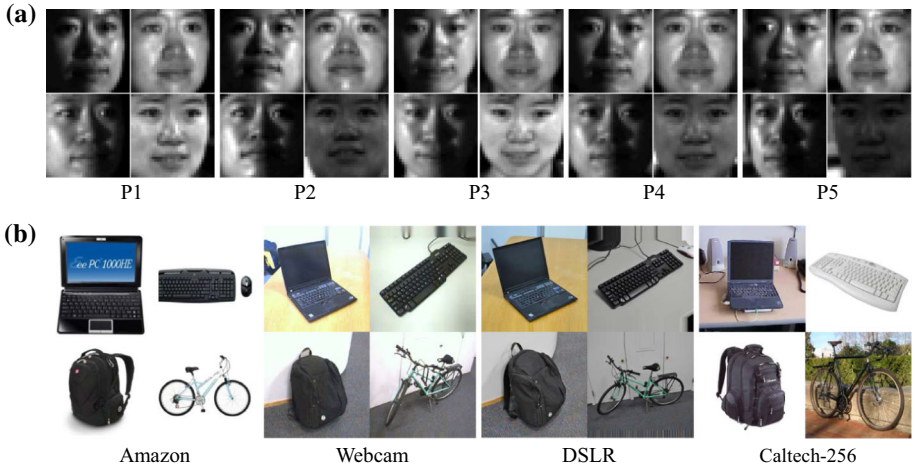

Fig. 2 Sample images from **a** PIE and **b** Office-Caltech datasets

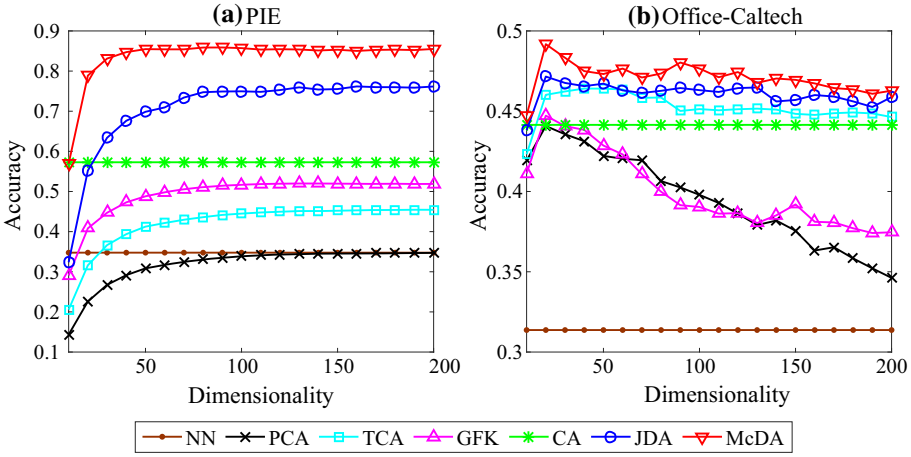

Fig. 3 Average accuracy versus reduced dimensionalities on **a** PIE and **b** Office-Caltech datasets

number of iterations N in McDA to 10 to guarantee convergence. For the compared subspace learning methods, the subspace dimensionality ranges from {10, 20, …, 200}. For TCA, JDA and McDA, we apply the linear transformation for the PIE dataset and the linear kernel for the Office-Caltech dataset as suggested by [20]. We adopt the broadly used classification accuracy on target dataset as the evaluation metric.

Figure 3 shows the compared classification accuracy versus different subspace dimensionalities. For two benchmark datasets, McDA is superior to all the compared methods in all dimensionalities in terms of average accuracy. The average accuracy is defined as the mean of classification accuracy over different cross-domain datasets.

Table 1 illustrates the accuracy of facial image classification on 20 cross-domain datasets of the PIE dataset. The results of the experiment show that our method outperforms the baseline methods in quantity. Importantly, McDA exceeds the average accuracy of JDA by 9.7%. Table 2 reports object recognition accuracy when the Office-Caltech dataset is used.

Table 1 Accuracy (%) on the PIE dataset

Dataset	NN	PCA	TCA	GFK	CA	JDA	McDA
P1 → P2	26.09	25.91	42.91	39.47	48.99	76.12	**87.17**
P1 → P3	26.59	26.10	42.59	47.86	51.65	71.94	**83.88**
P1 → P4	30.67	30.64	60.71	63.35	69.18	91.14	**95.55**
P1 → P5	16.67	16.73	30.09	37.50	41.48	50.86	**77.57**
P2 → P1	24.49	24.79	42.29	42.77	44.27	75.48	**85.41**
P2 → P3	46.63	46.38	51.90	56.56	58.15	80.64	**86.83**
P2 → P4	54.07	54.28	64.61	66.75	72.63	81.59	**91.53**
P2 → P5	26.53	26.23	34.07	40.69	40.56	63.36	**81.50**
P3 → P1	21.37	21.46	35.05	42.83	47.30	75.33	**88.12**
P3 → P2	41.01	41.07	47.45	56.05	54.02	79.01	**87.97**
P3 → P4	46.53	46.47	56.23	66.39	74.26	85.46	**87.74**
P3 → P5	26.23	26.16	33.33	47.12	50.92	66.30	**74.57**
P4 → P1	32.95	32.71	56.45	65.22	72.09	91.99	**95.50**
P4 → P2	62.68	62.74	68.45	75.63	78.76	90.91	**93.31**
P4 → P3	73.22	72.86	76.90	80.09	86.52	90.07	**90.75**
P4 → P5	37.19	37.13	41.73	53.86	63.24	73.22	**84.01**
P5 → P1	18.49	18.70	26.95	31.60	39.86	59.30	**76.71**
P5 → P2	24.19	24.19	31.74	35.67	42.48	70.96	**83.18**
P5 → P3	28.31	28.43	31.25	42.77	52.39	73.90	**82.48**
P5 → P4	31.24	31.15	34.27	47.67	56.83	75.94	**83.87**
Average	34.76	34.71	45.45	51.99	57.28	76.18	**85.88**

The best value of each row is highlighted in bold

Table 2 Accuracy (%) on the Office-Caltech dataset

Dataset	NN	PCA	TCA	GFK	CA	JDA	McDA
C → A	23.70	41.34	43.42	41.02	43.22	**44.15**	43.53
C → W	25.76	35.25	38.98	40.68	39.32	38.98	**44.41**
C → D	25.48	43.95	45.86	41.40	40.13	44.59	**50.96**
A → C	26.00	39.54	39.09	40.25	35.26	**41.50**	41.05
A → W	29.83	34.24	42.03	40.00	37.63	42.37	**44.41**
A → D	25.48	35.67	36.31	36.31	37.58	**45.22**	42.68
W → C	19.86	29.83	33.04	30.72	30.28	**35.26**	**35.26**
W → A	22.96	27.56	31.11	31.84	30.69	30.17	**37.37**
W → D	59.24	**93.63**	90.45	87.90	85.99	89.81	89.17
D → C	26.27	31.17	33.04	30.10	30.90	30.63	**34.82**
D → A	28.50	32.36	34.45	32.05	33.40	33.72	**36.64**
D → W	63.39	84.41	89.15	84.41	85.42	89.49	**89.83**
Average	31.37	44.08	46.41	44.72	44.15	47.16	**49.18**

The best value of each row is highlighted in bold

McDA also outweighs the baseline methods in most situations and its average accuracy is higher than those of the compared baseline methods. Based on the analysis above, the prominent performance of McDA implies the efficacy of the proposed MMCD.

5.3 Sensitivity Analysis

5.3.1 Kernel Choice

It is an open problem to choose suitable kernels for kernel-based learning methods. To investigate the effect of the kernels on the performance of McDA, we run McDA and the baseline JDA on two datasets by comparing various kernel and non-kernel based cases. For the kernel based cases, the linear kernel, the polynomial kernel of degree two $(x^T y + 1)^2$, the Gaussian kernel $\exp(-\|x - y\|_2^2 / 2\sigma^2)$, and the exponential kernel $\exp(-\|x - y\|_2 / \sigma)$ are used. For the non-kernel case, the linear transformation in the original space is performed. For both Gaussian and exponential kernels, σ is set to the median of pairwise distances between all the samples. For the PIE dataset, it is hard to perform the eigenvalue decomposition over the large kernel matrix for the sake of large sample sizes. A compromise solution is to construct a small dataset termed PIE-sub, which is a subset of the PIE dataset by randomly selecting 10 images from each class. As a result, each domain of PIE-sub consists of 680 images from 68 classes.

Table 3 shows that McDA consistently outperforms the baseline method JDA in all the cases. This implies that the MCD regularization term in McDA is beneficial for capturing more distribution information so as to promote the performance of domain adaptation in a variety of kernels. It is worth mentioning that in non-kernel condition the performance of McDA is also enhanced due to the presence of the MCD term.

5.3.2 Parameter Selection

Our McDA has two regularization parameters: λ which is to avoid the optimization problem to be ill-defined, and β which is to balance the importance of the MCD term. In order to analyze the effect of parameters on the performance, we vary λ and β from the range $\{10^{-3}, 10^{-2.5}, \ldots, 10^{2}\}$ and run McDA on PIE and Office-Caltech datasets. The reduced dimensionality is set to the optimal value according to Fig. 3, i.e., $k = 90$ for PIE and $k = 20$ for Office-Caltech. Figure 4a shows that small values of λ and relatively big values of β for the PIE dataset help achieve high accuracy. This could result from both significant mean and covariance differences between source and target data distributions. It can be seen from Fig. 4b that the

Table 3 Average accuracy (%) versus different kernels on PIE-sub and Office-Caltech datasets

	Linear	Polynomial	Gaussian	Exponential	Non-kernel	Average
PIE-sub						
JDA	65.59	62.95	44.56	34.63	68.96	55.34
McDA	**78.08**	**77.20**	**67.74**	**55.66**	**78.40**	**71.42**
Off.-Cal.						
JDA	47.16	46.15	45.45	44.23	46.96	45.99
McDA	**49.18**	**47.03**	**47.31**	**45.73**	**47.17**	**47.28**

The better value in each comparison group is highlighted in bold

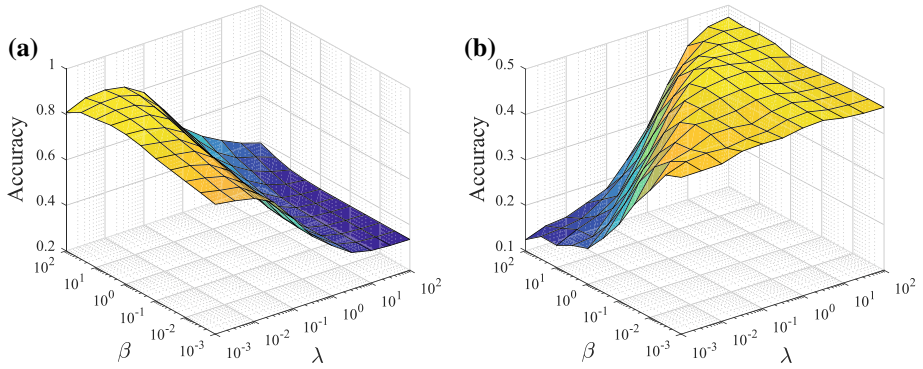

Fig. 4 Average accuracy of McDA versus different parameter values on **a** PIE and **b** Office-Caltech datasets

performance of McDA is relatively stable to the change of the values of λ and β in the area of $\lambda \geq \beta$ on the Office-Caltech dataset. Hence, $\lambda \in [0.001, 1]$ and $\beta \in [0.01, 10]$ are suggested for accomplishing enhanced performance.

6 Conclusion

This paper proposes a new distribution metric namely maximum mean and covariance discrepancy (MMCD) which unites MMD with the proposed maximum covariance discrepancy (MCD). MMCD is able to capture more information about distributions compared to MMD. Based on MMCD, we developed a new domain adaptation method in the joint distribution adaptation paradigm. Experiments conducted on two benchmark datasets verify the effectiveness of our method, which implies the efficacy of MMCD. In the future, we plan to apply the MMCD to other domain adaptation methods and test them on various applications.

Acknowledgements This work was supported by the National Natural Science Foundation of China (61806213, 61702134, U1435222).

Appendix A: Gradient Computation

According to (16), when the polynomial kernel of degree d is adopted, the gradient of the empirical estimator of squared MMD with respect to the data matrix $A = [X, Y]$ is given by

$$\frac{\partial \widehat{\mathrm{MMD}}^2}{\partial A} = 2dA(M \circ K_{d-1}), \tag{39}$$

where $(K_{d-1})_{ij} = \left(A_i^T A_j + c\right)^{d-1}$ and $\circ$ denotes the element-wise product. Likewise, there holds

$$\frac{\partial \widehat{\mathrm{MCD}}^2}{\partial A} = 4dA(ZK_dZ \circ K_{d-1}), \tag{40}$$

and

$$\frac{\partial \widehat{\mathrm{MMCD}}^2}{\partial A} = \frac{\partial \widehat{\mathrm{MMD}}^2}{\partial A} + \beta \frac{\partial \widehat{\mathrm{MCD}}^2}{\partial A}. \tag{41}$$

The gradients of MMD, MCD and MMCD with the linear kernel can be obtained by setting $d = 1$ and $c = 0$ in (39)–(41), respectively.

References

1. Baktashmotlagh M, Harandi M, Salzmann M (2016) Distribution-matching embedding for visual domain adaptation. J Mach Learn Res 17(1):3760–3789
2. Baktashmotlagh M, Harandi MT, Lovell BC, Salzmann M (2013) Unsupervised domain adaptation by domain invariant projection. In: IEEE international conference on computer vision, pp 769–776
3. Borgwardt KM, Gretton A, Rasch MJ, Kriegel HP, Schölkopf B, Smola AJ (2006) Integrating structured biological data by kernel maximum mean discrepancy. Bioinformatics 22(14):e49–e57
4. Bruzzone L, Marconcini M (2010) Domain adaptation problems: a DASVM classification technique and a circular validation strategy. IEEE Trans Pattern Anal Mach Intell 32(5):770–787
5. Cao X, Wipf D, Wen F, Duan G, Sun J (2013) A practical transfer learning algorithm for face verification. In: IEEE international conference on computer vision, pp 3208–3215
6. Dai W, Yang Q, Xue GR, Yu Y (2007) Boosting for transfer learning. In: Ghahramani Z (ed) Proceedings of the 24th international conference on machine learning. ACM, New York, pp 193–200
7. Dudley R, Fulton W, Katok A, Sarnak P, Bollobás B, Kirwan F (2002) Real analysis and probability. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511755347
8. Fernando B, Habrard A, Sebban M, Tuytelaars T (2013) Unsupervised visual domain adaptation using subspace alignment. In: IEEE International conference on computer vision, pp 2960–2967
9. Fukumizu K, Gretton A, Sun X, Schölkopf B (2008) Kernel measures of conditional dependence. In: Advances in neural information processing systems, pp 489–496
10. Gong B, Shi Y, Sha F, Grauman K (2012) Geodesic flow kernel for unsupervised domain adaptation. In: IEEE international conference on computer vision, pp 2066–2073
11. Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola A (2012) A kernel two-sample test. J Mach Learn Res 13(Mar):723–773
12. Gretton A, Bousquet O, Smola A, Schölkopf B (2005) Measuring statistical dependence with Hilbert–Schmidt norms. In: International conference on algorithmic learning theory, pp 63–77
13. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology
14. Guo Y, Ding G, Liu Q (2015) Distribution regularized nonnegative matrix factorization for transfer visual feature learning. In: ACM international conference on multimedia retrieval, pp 299–306
15. Hsieh YT, Tao SY, Tsai YHH, Yeh YR, Wang YCF (2016) Recognizing heterogeneous cross-domain data via generalized joint distribution adaptation. In: IEEE international conference on multimedia and expo, pp 1–6
16. Hsu TH, Chen W, Hou C, Tsai YH, Yeh Y, Wang YF (2015) Unsupervised domain adaptation with imbalanced cross-domain data. In: IEEE international conference on computer vision, pp 4121–4129
17. Huang J, Smola AJ, Gretton A, Borgwardt KM, Schölkopf B (2006) Correcting sample selection bias by unlabeled data. In: Advances in neural information processing systems, pp 601–608
18. Jiang W, Deng C, Liu W, Nie F, Chung F, Huang H (2017) Theoretic analysis and extremely easy algorithms for domain adaptive feature learning. In: Proceedings of the international joint conference on artificial intelligence, pp 1958–1964
19. Long M, Cao Y, Wang J, Jordan M (2015) Learning transferable features with deep adaptation networks. In: International conference on machine learning, pp 97–105
20. Long M, Wang J, Ding G, Sun J, Yu PS (2013) Transfer feature learning with joint distribution adaptation. In: IEEE international conference on computer vision, pp 2200–2207
21. Long M, Zhu H, Wang J, Jordan MI (2016) Unsupervised domain adaptation with residual transfer networks. In: Advances in neural information processing systems, pp 136–144
22. Mroueh Y, Sercu T, Goel V (2017) McGan: mean and covariance feature matching GAN. In: International conference on machine learning, pp 2527–2535
23. Muandet K, Fukumizu K, Sriperumbudur B, Schölkopf B et al (2017) Kernel mean embedding of distributions: a review and beyond. Found Trends Mach Learn 10(1–2):1–141

24. Müller A (1997) Integral probability metrics and their generating classes of functions. Adv Appl Probab 29(2):429–443
25. Pan SJ, Tsang IW, Kwok JT, Yang Q (2011) Domain adaptation via transfer component analysis. IEEE Trans Neural Netw 22(2):199–210
26. Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
27. Patel VM, Gopalan R, Li R, Chellappa R (2015) Visual domain adaptation: a survey of recent advances. IEEE Signal Process Mag 32(3):53–69
28. Quang M.H, San Biagio M, Murino V (2014) Log-Hilbert–Schmidt metric between positive definite operators on Hilbert spaces. In: Advances in neural information processing systems, pp 388–396
29. Quang Minh H, San Biagio M, Bazzani L, Murino V (2016) Approximate log-Hilbert–Schmidt distances between covariance operators for image classification. In: IEEE conference on computer vision and pattern recognition, pp 5195–5203
30. Saenko K, Kulis B, Fritz M, Darrell T (2010) Adapting visual category models to new domains. In: European conference on computer vision, pp 213–226
31. Schölkopf B, Smola AJ, Bach F et al (2002) Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT Press, Cambridge
32. Shen J, Qu Y, Zhang W, Yu Y (2018) Wasserstein distance guided representation learning for domain adaptation. In: AAAI conference on artificial intelligence
33. Si S, Tao D, Geng B (2010) Bregman divergence-based regularization for transfer subspace learning. IEEE Trans Knowl Data Eng 22(7):929–942
34. Sriperumbudur BK, Gretton A, Fukumizu K, Lanckriet GRG, Schölkopf B (2008) Injective Hilbert space embeddings of probability measures. In: Annual conference on learning theory, pp 111–122
35. Sun B, Feng J, Saenko K (2016) Return of frustratingly easy domain adaptation. In: AAAI conference on artificial intelligence, pp 2058–2065
36. Sun H, Liu S, Zhou S (2016) Discriminative subspace alignment for unsupervised visual domain adaptation. Neural Process Lett 44(3):779–793
37. Wang T, Ye T, Gurrin C (2016) Transfer nonnegative matrix factorization for image representation. In: International conference on multimedia modeling, pp 3–14
38. Xie X, Sun S, Chen H, Qian J (2018) Domain adaptation with twin support vector machines. Neural Process Lett 48(2):1213–1226
39. Zellinger W, Grubinger T, Lughofer E, Natschläger T, Saminger-Platz S (2017) Central moment discrepancy (CMD) for domain-invariant representation learning. In: International conference on learning representations
40. Zhu L, Zhang X, Zhang W, Huang X, Guan N, Luo Z (2017) Unsupervised domain adaptation with joint supervised sparse coding and discriminative regularization term. In: IEEE international conference on image processing, pp 3066–3070
41. Zong Y, Zheng W, Zhang T, Huang X (2016) Cross-corpus speech emotion recognition based on domain-adaptive least-squares regression. IEEE Signal Process Lett 23(5):585–589