Manifold Embedded Knowledge Transfer for Brain-Computer Interfaces

Wen Zhang and Dongrui Wu, *Senior Member, IEEE*

***Abstract*— Transfer learning makes use of data or knowledge in one problem to help solve a different, yet related, problem. It is particularly useful in brain-computer interfaces (BCIs), for coping with variations among different subjects and/or tasks. This paper considers offline unsupervised cross-subject electroencephalogram (EEG) classification, i.e., we have labeled EEG trials from one or more source subjects, but only unlabeled EEG trials from the target subject. We propose a novel manifold embedded knowledge transfer (MEKT) approach, which first aligns the covariance matrices of the EEG trials in the Riemannian manifold, extracts features in the tangent space, and then performs domain adaptation by minimizing the joint probability distribution shift between the source and the target domains, while preserving their geometric structures. MEKT can cope with one or multiple source domains, and can be computed efficiently. We also propose a domain transferability estimation (DTE) approach to identify the most beneficial source domains, in case there are a large number of source domains. Experiments on four EEG datasets from two different BCI paradigms demonstrated that MEKT outperformed several state-of-the-art transfer learning approaches, and DTE can reduce more than half of the computational cost when the number of source subjects is large, with little sacrifice of classification accuracy.**

***Index Terms*— Brain-computer interfaces, electroencephalogram, Riemannian manifold, transfer learning.**

I. Introduction

A BRAIN-COMPUTER interface (BCI) provides a direct communication pathway between a user's brain and a computer [1], [2]. Electroencephalogram (EEG), a multi-channel time-series, is the most frequently used BCI input signal. There are three common paradigms in EEG-based BCIs: motor imagery (MI) [3], event-related potentials (ERPs) [4], and steady-state visual evoked potentials [2]. The first two are the focus of this paper.

In MI tasks, the user needs to imagine the movements of his/her body parts, which causes modulations of brain rhythms in the involved cortical areas. In ERP tasks, the user is stimulated by a majority of non-target stimuli and a few target stimuli; a special ERP pattern appears in the EEG response after the user perceives a target stimulus. EEG-based BCI systems have been widely used to help people with disabilities, and also the able-bodied [1].

A standard EEG signal analysis pipeline consists of temporal (band-pass) filtering, spatial filtering, and classification [5]. Spatial filters such as common spatial patterns (CSP) [6] are widely used to enhance the signal-to-noise ratio. Recently, there is a trend to utilize the covariance matrices of EEG trials, which are symmetric positive definite (SPD) and can be viewed as points on a Riemannian manifold, in EEG signal analysis [7]–[10]. For MI tasks, the discriminative information is mainly spatial, and can be directly encoded in the covariance matrices. On the contrary, the main discriminative information of ERP trials is temporal. A novel approach was proposed in [11] to augment each EEG trial by the mean of all target trials that contain the ERP, and then their covariance matrices are computed. However, Riemannian space based approaches are computationally expensive, and not compatible with Euclidean space machine learning approaches.

A major challenge in BCIs is that different users have different neural responses to the same stimulus, and even the same user can have different neural responses to the same stimulus at different time/locations. Besides, when calibrating the BCI system, acquiring a large number of subject-specific labeled training examples for each new subject is time-consuming and expensive. Transfer learning [12]–[16], which uses data/information from one or more source domains to help the learning in a target domain, can be used to address these problems. Some representative applications of transfer learning in BCIs can be found in [17]–[22]. Many researchers [20]–[22] attempted to seek a set of subject-invariant CSP filters to increase the signal-to-noise ratio. Another pipeline is Riemannian geometry based. Zanini *et al.* [23] proposed a Riemannian alignment (RA) framework to align the EEG covariance matrices from different subjects. He and Wu [24] extended RA to Euclidean alignment (EA) in the Euclidean space, so that any Euclidean space classifier can be used after it.

To utilize the excellent properties of the Riemannian geometry and avoid its high computational cost, as well as to leverage knowledge learned from the source subjects, this paper proposes a manifold embedded knowledge transfer (MEKT) framework, which first aligns the covariance matrices of the EEG trials in the Riemannian manifold, then performs domain adaptation in the tangent space by minimizing the joint probability distribution shift between the source and the target domains, while preserving their geometric

Manuscript received October 9, 2019; revised February 29, 2020; accepted April 2, 2020. Date of publication April 6, 2020; date of current version May 8, 2020. This work was supported in part by the Hubei Technology Innovation Platform under Grant 2019AEA171 and in part by the National Natural Science Foundation of China under Grant 61873321. *(Corresponding author: Dongrui Wu.)*

The authors are with the Ministry of Education Key Laboratory of Image Processing and Intelligent Control, School of Artificial Intelligence and Automation, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: wenz@hust.edu.cn; drwu@hust.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2020.2985996

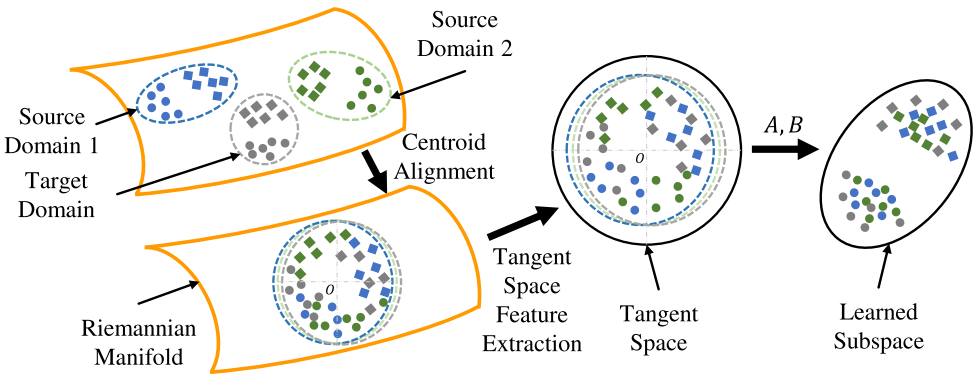

Fig. 1. Illustration of our proposed MEKT. Squares and circles represent examples from different classes. Different colors represent different domains. All domains are first aligned on the Riemannian manifold, and then mapped onto a tangent space. *A* and *B* are projection matrices of the source and the target domains, respectively.

structures, as illustrated in Fig. 1. Additionally, we propose a domain transferability estimation (DTE) approach to select the most beneficial subjects in multi-source transfer learning. Experiments on four datasets from two different BCI paradigms (MI and ERP) verified the effectiveness of MEKT and DTE.

The remainder of this paper is organized as follows: Section II introduces related work on spatial filters, Riemannian geometry, tangent space mapping, RA, EA, and subspace adaptation. Section III describes the details of the proposed MEKT and DTE. Section IV presents experiments to compare the performance of MEKT with several state-of-the-art data alignment and transfer learning approaches. Finally, Section V draws conclusions.

II. Related Work

This section introduces background knowledge on spatial filters, Riemannian geometry, tangent space mapping, RA, EA, and subspace adaptation, which will be used in the next section.

A. Spatial Filters

Spatial filtering can be viewed as a data-driven dimensionality reduction approach that promotes the variance difference between two conditions [25]. It is common in MI-based BCIs to use CSP filters [26] to simultaneously diagonalize the two intra-class covariance matrices.

Consider a binary classification problem. Let (X_i, y_i) be the ith labeled training example, where $X_i \in \mathbb{R}^{c\times t}$, in which c is the number of EEG channels, and t the number of time domain samples. For Class k ($k = 0, 1$), CSP finds a spatial filter matrix $W_k^* \in \mathbb{R}^{c\times f}$, where f is the number of spatial filters, to maximize the variance difference between Class k and Class $1-k$:

$$W_k^* = \arg\max_{W\in\mathbb{R}^{c\times f}} \frac{\mathrm{tr}(W^\top \bar{\Sigma}_k W)}{\mathrm{tr}[W^\top \bar{\Sigma}_{1-k} W]}, \tag{1}$$

where $\bar{\Sigma}_k$ is the mean covariance matrix of all EEG trials in Class k, and tr is the trace of a matrix. The solution W_k^* is the concatenation of the f leading eigenvectors of the matrix $(\bar{\Sigma}_{1-k})^{-1}\bar{\Sigma}_k$.

Finally, we concatenate the $2f$ spatial filters from both classes to obtain the complete CSP filters:

$$W^* = [W_0^*, W_1^*] \in \mathbb{R}^{c\times 2f} \tag{2}$$

and compute the spatially filtered X_i by:

$$X_i' = (W^*)^\top X_i \in \mathbb{R}^{2f\times t} \tag{3}$$

The log-variances of the filtered trial X' can be extracted:

$$\mathbf{x} = \log\left(\frac{\mathrm{diag}(X'X'^\top)}{\mathrm{tr}(X'X'^\top)}\right) \tag{4}$$

and used as input features in classification.

B. Riemannian Geometry

All SPD matrices $P \in \mathbb{R}^{c\times c}$ form a differentiable Riemannian manifold. Riemannian geometry is used to manipulate them. Some basic definitions are provided below.

The Riemannian distance between two SPD matrices P_1 and P_2 is:

$$\delta(P_1, P_2) = \left\|\log\left(P_1^{-1}P_2\right)\right\|_F, \tag{5}$$

where $\|\cdot\|_F$ is the Frobenius norm, and log donates the logarithm of the eigenvalues of $P_1^{-1}P_2$.

The Riemannian mean of $\{P_i\}_{i=1}^n$ is:

$$M_R = \arg\min_P \sum_{i=1}^n \delta^2(P, P_i), \tag{6}$$

The Euclidean mean is:

$$M_E = \frac{1}{n}\sum_{i=1}^n P_i, \tag{7}$$

The Log-Euclidean mean [7] is:

$$M_L = \exp\left(\sum_{i=1}^n w_i \log P_i\right), \tag{8}$$

where w_i is usually set to $\frac{1}{n}$.

C. Tangent Space Mapping

Tangent space mapping is also known as the logarithmic mapping, which maps a Riemannian space SPD matrix P_i to a Euclidean tangent space vector $\mathbf{x}_i$ around an SPD matrix M, which is usually the Riemannian or Euclidean mean:

$$\mathbf{x}_i = \text{upper}\left(\log_M\left(M_{\text{ref}} P_i M_{\text{ref}}\right)\right), \tag{9}$$

where upper takes the upper triangular part of a $c \times c$ SPD matrix and forms a vector $\mathbf{x}_i \in \mathbb{R}^{1\times c(c+1)/2}$, and M_{ref} is a reference matrix. To obtain a tangent space locally homomorphic to the manifold, $M_{\text{ref}} = M^{-1/2}$ is needed [25].

Congruent transform and congruence invariance [27] are two important properties in the Riemannian space:

$$\mathcal{M}\left(FP_1\ F, FP_2\ F\right) = F \cdot \mathcal{M}(P_1, P_2) \cdot F, \tag{10}$$

$$\delta\left(G^\top P_1 G, G^\top P_2 G\right) = \delta\left(P_1, P_2\right), \tag{11}$$

where $\mathcal{M}$ is the Euclidean or Riemannian mean operation, F is a nonsingular square matrix, and $G \in \mathbb{R}^{c\times c}$ is an invertible symmetric matrix. (11) suggests that the Riemannian distance between two SPD matrices does not change, if both are left and right multiplied by an invertible symmetric matrix.

D. Riemannian Alignment (RA)

RA [23] first computes the covariance matrices of some resting (or non-target) trials, $\{P_i\}_{i=1}^n$, in which the subject is not performing any task (or not performing the target task), and then the Riemannian mean M_R of these matrices, which is used as the reference matrix to reduce the inter-session or inter-subject variations, by the following transformation:

$$P_i' = M_R^{-1/2} P_i M_R^{-1/2}, \tag{12}$$

where P_i is the covariance matrix of the i-th trial, and P_i' the corresponding aligned covariance matrix. Then, all P_i' can be classified by a minimum distance to mean (MDM) classifier [8].

E. Euclidean Alignment (EA)

Although RA-MDM has demonstrated promising performance, it still has some limitations [24]: 1) it processes the covariance matrices in the Riemannian space, whereas there are very few Riemannian space classifiers; 2) it computes the reference matrix from the non-target stimuli in ERP-based BCIs, which requires some labeled trials from the new subject.

EA [24] extends RA and solves the above problems by transforming an EEG trial X_i in the Euclidean space:

$$X_i' = M_E^{-1/2} X_i, \tag{13}$$

where M_E is the Euclidean mean of the covariance matrices of all EEG trials, computed by (7).

However, EA only considers the marginal probability distribution shift, and works best when the number of EEG channels is small. When there are a large number of channels, computing $M_E^{-1/2}$ may be numerically unstable.

F. Subspace Adaptation

Tangent space vectors usually have very high dimensionality, so they cannot be used easily in transfer learning. An intuitive approach is to align them in a lower dimensional subspace. Pan *et al.* [12] proposed transfer component analysis (TCA) to learn the transferable components across domains in a reproducible kernel Hilbert space using maximum mean discrepancy (MMD) [28]. Joint distribution adaptation (JDA) [15] improves TCA by considering the conditional distribution shift using pseudo label refinement. Joint geometrical and statistical alignment (JGSA) [16] further improves JDA by adding two regularization terms, which minimize the within-class scatter matrix and maximize the between-class scatter matrix.

III. Manifold Embedded Knowledge Transfer (MEKT)

This section proposes the MEKT approach. Its goal is to use one or multiple source subjects' data to help the target subject, given that they have the same feature space and label space. For the ease of illustration, we focus on a single source domain first.

Assume the source domain has n_S labeled instances $\{(X_{S,i}, y_{S,i})\}_{i=1}^{n_S}$, where $X_{S,i} \in \mathbb{R}^{c\times t}$ is the i-th feature matrix, and $y_{S,i} \in \{1, \ldots, l\}$ is the corresponding label, in which c, t and l denote the number of channels, time domain samples, and classes, respectively. Let $\mathbf{y}_S = [y_{S,1}; \cdots; y_{S,n_S}] \in \mathbb{R}^{n_S\times 1}$ be the label vector of the source domain. Assume also the target domain has n_T unlabeled feature matrices $\{X_{T,i}\}_{i=1}^{n_T}$, where $X_{T,i} \in \mathbb{R}^{c\times t}$.

MEKT consists of the following three steps:

1) *Covariance matrix centroid alignment (CA)*: Align the centroids of the covariance matrices of $\{X_{S,i}\}_{i=1}^{n_S}$ and $\{X_{T,i}\}_{i=1}^{n_T}$, so that their marginal probability distributions are close.
2) *Tangent space feature extraction*: Map the aligned covariance matrices to a tangent space feature matrix $X_S \in \mathbb{R}^{d\times n_S}$, and $X_T \in \mathbb{R}^{d\times n_T}$, where $d = c(c+1)/2$ is the dimensionality of the tangent space features.
3) *Mapping matrices identification*: Find projection matrices $A \in \mathbb{R}^{d\times p}$ and $B \in \mathbb{R}^{d\times p}$, where $p \ll d$ is the dimensionality of a shared subspace, such that $A^\top X_S$ and $B^\top X_T$ are close.

After MEKT, a classifier can be trained on $(A^\top X_S, \mathbf{y}_S)$ and applied to $B^\top X_T$ to obtain the target pseudo labels $\hat{\mathbf{y}}_T$.

Next, we describe the details of the above three steps.

A. Covariance Matrix Centroid Alignment (CA)

CA serves as a preprocessing step to reduce the marginal probability distribution shift of different domains, and enables transfer from multiple source domains.

Let $P_{S,i} = X_{S,i} X_{S,i}^\top$ be the i-th covariance matrix in the source domain, and $M_{\text{ref}} = M^{-1/2}$, where M can be the Riemannian mean in (6), the Euclidean mean in (7), or the Log-Euclidean mean in (8). Then, we align the covariance matrices by

$$P_{S,i}' = M_{\text{ref}} P_{S,i} M_{\text{ref}}, \qquad i = 1, \ldots, n_S \tag{14}$$

Likewise, we can obtain the aligned covariance matrices $\{P'_{T,i}\}_{i=1}^{n_T}$ of the target domain.

CA has two desirable properties:

1) *Marginal probability distribution shift minimization*. From the properties of congruent transform and congruence invariance, we have

$$\mathcal{M}(M_{\text{ref}}^{\top} P_1 M_{\text{ref}}, \ldots, M_{\text{ref}}^{\top} P_{n_S} M_{\text{ref}}) = M_{\text{ref}}^{\top} \mathcal{M}(P_1, \ldots, P_{n_S}) M_{\text{ref}} = M_{\text{ref}}^{\top} M M_{\text{ref}} = I, \quad (15)$$

i.e., if we choose M as the Riemannian (or Euclidean) mean, then different domains' geometric (or arithmetic) centers all equal the identity matrix. Therefore, the marginal distributions of the source and the target domains are brought closer on the manifold.

2) *EEG trial whitening*. In the following, we show that each aligned covariance matrix is approximately an identity matrix after CA.

If we decompose the reference matrix as $M_{\text{ref}} = [\mathbf{w}_1, \ldots, \mathbf{w}_c]$, then the (m, n)-th element of $P'_{S,i}$ is:

$$P'_{S,i}(m, n) = \mathbf{w}_m^{\top} P_{S,i} \mathbf{w}_n, \quad (16)$$

From (15) we have

$$\mathbf{w}_m^{\top} \mathcal{M}(P_1, \ldots, P_{n_S}) \mathbf{w}_n = \begin{cases} 1, & m = n \\ 0, & m \neq n. \end{cases} \quad (17)$$

The above equation holds no matter whether $\mathcal{M}$ is the Riemannian mean, or the Euclidean mean.

For CA using the Euclidean mean, the average of the m-th diagonal element of $\{P'_{S,i}\}_{i=1}^{n_S}$ is

$$\frac{1}{n_S} \sum_{i=1}^{n_S} P'_{S,i}(m, m) = \mathbf{w}_m^{\top} \mathcal{M}(P_1, \ldots, P_{n_S}) \mathbf{w}_m = 1, \quad (18)$$

Meanwhile, for each diagonal element, we have $P'_{S,i}(m, m) = \|X_{S,i}^{\top} \mathbf{w}_m\|_2^2 > 0$, therefore the diagonal elements of $P'_{S,i}$ are around 1. Similarly, the off-diagonal elements of $P'_{S,i}$ are around 0. Thus, $P'_{S,i}$ is approximatively an identify matrix, i.e., the aligned EEG trials are approximated whitened.

CA with the Riemannian mean is an iterative process initialized by the Euclidean mean. CA with the Log-Euclidean mean is an approximation of CA with the Riemannian mean, with reduced computational cost [7]. So, (18) also holds approximately for these two means. This whitening effect will also be experimentally demonstrated in Section IV-E.

B. Tangent Space Feature Extraction

After covariance matrix CA, we map each covariance matrix to a tangent space feature vector in $\mathbb{R}^{d \times 1}$:

$$\mathbf{x}_{S,i} = \text{upper}\left(\log\left(P'_{S,i}\right)\right), \quad i = 1, \ldots, n_S \quad (19)$$

$$\mathbf{x}_{T,i} = \text{upper}\left(\log\left(P'_{T,i}\right)\right), \quad i = 1, \ldots, n_T \quad (20)$$

Note that this is different from the original tangent space mapping in (9), in that (9) uses the same reference matrix M_{ref} for all subjects, whereas our approach uses a subject-specific M_{ref} for each different subject.

Next, we form new feature matrices $X_S = [\mathbf{x}_{S,i}, \ldots, \mathbf{x}_{S,n_S}]$ and $X_T = [\mathbf{x}_{T,i}, \ldots, \mathbf{x}_{T,n_T}]$.

C. Mapping Matrices Identification

CA does not reduce the conditional probability distribution discrepancies. We next find projection matrices $A, B \in \mathbb{R}^{d \times d'}$, which map X_S and X_T to lower dimensional matrices $A^{\top} X_S$ and $B^{\top} X_T$, with the following desirable properties:

1) *Joint probability distribution shift minimization.* In traditional domain adaptation [12], [15], MMD is frequently used to reduce the marginal and conditional probability distribution discrepancies between the source and the target domains, i.e.,

$$\begin{aligned} \mathcal{D}_{S,T} &\approx \mathcal{D}\left(Q\left(X_S\right), Q\left(X_T\right)\right) + \mathcal{D}\left(Q\left(\mathbf{y}_S|X_S\right), Q\left(\hat{\mathbf{y}}_T|X_T\right)\right) \\ &= \left\| \frac{1}{n_S} \sum_{i=1}^{n_S} A^{\top} \mathbf{x}_{S,i} - \frac{1}{n_T} \sum_{j=1}^{n_T} B^{\top} \mathbf{x}_{T,j} \right\|_F^2 \\ &\quad + \sum_{k=1}^{l} \left\| \frac{1}{n_S^k} \sum_{i=1}^{n_S^k} A^{\top} \mathbf{x}_{S,i}^k - \frac{1}{n_T^k} \sum_{j=1}^{n_T^k} B^{\top} \mathbf{x}_{T,j}^k \right\|_F^2, \end{aligned} \quad (21)$$

where $\mathbf{x}_{S,i}^k$ and $\mathbf{x}_{T,j}^k$ are the tangent space vectors in the k-th ($k = 1, \ldots, l$) class of the source domain and the target domain, respectively, and n_S^k and n_T^k are the number of examples in the k-th class of the source domain and the target domain, respectively.

Next, we propose a new measure, joint probability MMD, to quantify the probability distribution shift between the source and the target domains, by considering the joint probability directly, instead of the marginal and the conditional probabilities separately.

Then, the joint probability MMD between the source and the target domains is:

$$\begin{aligned} \mathcal{D}'_{S,T} &= \mathcal{D}\left(Q\left(X_S, \mathbf{y}_S\right), Q\left(X_T, \hat{\mathbf{y}}_T\right)\right) \\ &= \mathcal{D}\left(Q\left(X_S|\mathbf{y}_S\right) Q(\mathbf{y}_S), Q\left(X_T|\hat{\mathbf{y}}_T\right) Q(\hat{\mathbf{y}}_T)\right) \\ &\approx \sum_{k=1}^{l} \left\| \frac{P(\mathbf{y}_S^k)}{n_S^k} \sum_{i=1}^{n_S^k} A^{\top} \mathbf{x}_{S,i}^k - \frac{P(\hat{\mathbf{y}}_T^k)}{n_T^k} \sum_{j=1}^{n_T^k} B^{\top} \mathbf{x}_{T,j}^k \right\|_F^2 \\ &= \sum_{k=1}^{l} \left\| \frac{1}{n_S} \sum_{i=1}^{n_S^k} A^{\top} \mathbf{x}_{S,i}^k - \frac{1}{n_T} \sum_{j=1}^{n_T^k} B^{\top} \mathbf{x}_{T,j}^k \right\|_F^2, \end{aligned} \quad (22)$$

Let the one-hot encoding matrix of the source domain label vector[1] $\mathbf{y}_S$ be Y_S, and the one-hot encoding matrix of the *predicted* target label vector $\hat{\mathbf{y}}_T$ be $\hat{Y}_T$. (22) can

[1]For example, for binary classification with two classes 1 and 2, if $\mathbf{y}_S = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$, then $Y_S = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$.

be simplified as

$$\mathcal{D}'_{S,T} = \left\| N_S^\top X_S^\top A - N_T^\top X_T^\top B \right\|_F^2, \tag{23}$$

where

$$N_S = \frac{Y_S}{n_S}, \quad N_T = \frac{\hat{Y}_T}{n_T}. \tag{24}$$

The joint probability MMD is based on the joint probability rather than the conditional probability, which in theory can handle more probability distribution shifts.

2) *Source domain discriminability.* During subspace mapping, the discriminating ability of the source domain can be preserved by:

$$\min_A \ \mathrm{tr}(A^\top S_w A) \quad s.t.\ A^\top S_b A = I, \tag{25}$$

where $S_w = \sum_{k=1}^{l} \sum_{i=1}^{n_S^k} (\mathbf{x}_{S,i}^k)^\top \mathbf{x}_{S,i}^k h_k$ is the within-class scatter matrix, in which $h_k = 1 - \frac{1}{n_S^k}$, and $S_b = \sum_{k=1}^{l} n_k (\bar{m}_k - \bar{m})(\bar{m}_k - \bar{m})^\top$ is the between-class scatter matrix, in which $\bar{m}_k$ is the mean of samples from Class k, and $\bar{m}$ is the mean of all samples.

3) *Target domain locality preservation.* We also introduce a graph-based regularization term to preserve the local structure in the target domain. Under the manifold assumption [29], if two samples $\mathbf{x}_{T,i}$ and $\mathbf{x}_{T,j}$ are close in the original target domain, then they should also be close in the projected subspace.

Let $S \in \mathbb{R}^{n_T \times n_T}$ be a similarity matrix:

$$S_{ij} = \begin{cases} e^{\frac{-\|\mathbf{x}_{T,i} - \mathbf{x}_{T,j}\|_2^2}{2\sigma^2}}, & \text{if } \mathbf{x}_{T,i} \in N_p(\mathbf{x}_{T,j}) \\ & \text{or } \mathbf{x}_{T,j} \in N_p(\mathbf{x}_{T,i}) \\ 0, & \text{otherwise,} \end{cases} \tag{26}$$

where $N_p(\mathbf{x}_{T,i})$ is the set of the p-nearest neighbors of $\mathbf{x}_{T,i}$, and σ is a scaling parameter, which usually equals 1 [30].

Using the normalized graph Laplacian matrix $L = I - D^{-1/2} S D^{-1/2}$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n_T} S_{ij}$, graph regularization is expressed as:

$$\sum_{i,j=1}^{n_T} \|B^\top \mathbf{x}_{T,i} - B^\top \mathbf{x}_{T,j}\|_2^2 S_{ij} = \mathrm{tr}(B^\top X_T L X_T^\top B), \tag{27}$$

To remove the scaling effect, we add a constraint on the target embedding [31]:

$$\min_B \mathrm{tr}(B^\top X_T L X_T^\top B) \quad s.t.\ B^\top X_T H X_T^\top B = I, \tag{28}$$

where $H = I - \frac{1}{n_T} \mathbf{1}_{n_T}$ is the centering matrix, in which $\mathbf{1}_{n_T} \in \mathbb{R}^{n_T \times n_T}$ is an all-one matrix.

4) *Parameter transfer and regularization.* Since the source and the target domains have the same feature space, and CA has brought their probability distributions closer, we want the projection matrix B to be similar to the projection matrix A learned in the source domain. Additionally, for better generalization performance, we want to ensure that A and B do not include extreme values. Thus, we have the following constraints on the projection matrices:

$$\min_{A,B} \left(\|B - A\|_F^2 + \|B\|_F^2 \right). \tag{29}$$

D. The Overall Loss Function of MEKT

Integrating all regularization and constraints above, the formulation of MEKT is:

$$\begin{aligned} \min_{A,B} \ & \alpha\, \mathrm{tr}(A^\top S_w A) + \beta\, \mathrm{tr}(B^\top X_T L X_T^\top B) + \mathcal{D}'_{S,T} \\ & + \rho(\|B - A\|_F^2 + \|B\|_F^2) \\ s.t.\ & B^\top X_T H X_T^\top B = I, \quad A^\top S_b A = I \end{aligned} \tag{30}$$

where α, β and ρ are trade-off parameters to balance the importance of the source domain discriminability, the target domain locality, and the parameter regularization, respectively.

Let $W = [A; B]$. Then, the Lagrange function is

$$\mathcal{J} = \mathrm{tr}\left(W^\top (\alpha P + \beta L + \rho U + R) W + \eta (I - W^\top V W) \right) \tag{31}$$

where

$$P = \begin{bmatrix} S_w & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}, \quad L = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & X_T L X_T^\top \end{bmatrix}, \tag{32}$$

$$U = \begin{bmatrix} I & -I \\ -I & 2I \end{bmatrix}, \quad V = \begin{bmatrix} S_b & \mathbf{0} \\ \mathbf{0} & X_T H X_T^\top \end{bmatrix}, \tag{33}$$

$$R = \begin{bmatrix} X_S N_S N_S^\top X_S^\top & -X_S N_S N_T^\top X_T^\top \\ -X_T N_T N_S^\top X_S^\top & X_T N_T N_T^\top X_T^\top \end{bmatrix}, \tag{34}$$

Setting the derivative $\nabla_W \mathcal{J} = \mathbf{0}$, we have

$$(\alpha P + \beta L + \rho U + R) W = \eta V W \tag{35}$$

(35) can be solved by generalized eigen-decomposition, and W consists of the p trailing eigenvectors. Since $\hat{Y}_T$ is needed in N_T [see (24)], and hence R, we use a general expectation-maximization like pseudo label refinement procedure [15] to refine the estimation, as shown in Algorithm 1.

Note that for the clarity of explanation, Algorithm 1 only considers one source domain. When there are multiple source domains, we perform CA and compute the tangent space feature vectors $X_S^{(i)} \in \mathbb{R}^{d \times n_S^{(i)}}$ for each source domain separately, and then assemble their feature vectors into a single source domain feature matrix $X_S = [X_S^{(1)}, \ldots, X_S^{(z)}] \in \mathbb{R}^{d \times n^*}$, where $n_S^{(i)}$ is the number of trails in the i-th source domain, z is the number of source domains, and $n^* = \sum_{i=1}^{z} n_S^{(i)}$.

E. Kernelization Analysis

Nonlinear MEKT can be achieved through kernelization in a Reproducing Kernel Hilbert Space [16].

Let the kernel function be $\phi : \mathbf{x} \mapsto \phi(\mathbf{x})$. Define $\Phi(X) = [\phi(\mathbf{x}_1), .., \phi(\mathbf{x}_n)] \in \mathbb{R}^{d \times n}$, where $n = n_S + n_T$. We use the Representer Theorem [32] $A = \Phi(X)\mathbf{A}$ and $B = \Phi(X)\mathbf{B}$ to

Algorithm 1: Manifold Embedded Knowledge Transfer (MEKT)

Input: n_S source domain samples $\{(X_{S,i}, y_{S,i})\}_{i=1}^{n_S}$, where $X_{S,i} \in \mathbb{R}^{c\times t}$ and $y_{S,i} \in \{1, \ldots, l\}$; n_T target domain feature matrices $\{X_{T,i}\}_{i=1}^{n_T}$, where $X_{T,i} \in \mathbb{R}^{c\times t}$; Number of iterations N; Weights α, β, ρ; Dimensionality of the shared subspace, p.
Output: $\hat{\mathbf{y}}_T \in \mathbb{R}^{n_T\times 1}$, the labels for $\{X_{T,i}\}_{i=1}^{n_T}$.
Calculate the covariance matrices $\{P_{S,i}\}_{i=1}^{n_S}$ and their mean matrix M in the source domain, using (6), (7), or (8);
Calculate $\{P'_{S,i}\}_{i=1}^{n_S}$ using (14);
Map each $P'_{S,i}$ to a tangent space feature vector $\mathbf{x}_{S,i} \in \mathbb{R}^{d\times 1}$ using (19) ($d = c(c+1)/2$);
Repeat the above procedure to get $\mathbf{x}_{T,i} \in \mathbb{R}^{d\times 1}$ using (20);
Form $X_S = [\mathbf{x}_{S,1}, \ldots, \mathbf{x}_{S,n_S}]$ and $X_T = [\mathbf{x}_{T,1}, \ldots, \mathbf{x}_{T,n_T}]$;
Construct P, L, U, V and R in (32)-(34);
for $n = 1, \ldots, N$ **do**
  Solve (35), and construct $W \in \mathbb{R}^{2d\times p}$ as the p trailing eigenvectors;
  Construct A as the first d rows in W, and B as the last d rows;
  Train a classifier f on $(A^\top X_S, \mathbf{y}_S)$ and apply it to $B^\top X_T$ to update $\hat{\mathbf{y}}_T$;
  Update R in (34).
end
return $\hat{\mathbf{y}}_T$.

kernelize MEKT, where $X = [X_S, X_T]$, and $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$ are two projection matrices to be optimized.

Let $K_S = \Phi(X)^\top \Phi(X_S)$ and $K_T = \Phi(X)^\top \Phi(X_T)$. Then, all $\mathbf{x}$ are replaced by $\phi(\mathbf{x})$, X_S by $\Phi(X_S)$, and X_T by $\Phi(X_T)$, in the above derivations. The optimization problem becomes

$$\begin{aligned}\min_{\mathbf{A},\mathbf{B}}\ & \alpha\,\mathrm{tr}(\mathbf{A}^\top S_w \mathbf{A}) + \beta\,\mathrm{tr}(\mathbf{B}^\top K_T L K_T^\top \mathbf{B}) \\ & + \left\| N_S^\top K_S^\top \mathbf{A} - N_T^\top K_T^\top \mathbf{B} \right\|_F^2 + \rho(W^\top \mathbf{U} W) \\ s.t.\ & \mathbf{B}^\top K_T H K_T^\top \mathbf{B} = I, \quad \mathbf{A}^\top S_b \mathbf{A} = I \end{aligned} \tag{36}$$

where $S_w = \sum_{k=1}^{l} K_S^k H_S^k (K_S^k)^\top$, in which K_S^k is the part of K_S from Class k only, and $H_S^k = I - \frac{1}{n_S^k}\mathbf{1}$ the centering matrix. The Laplacian matrix L is constructed in the original data space. In S_b, $\bar{m}_k$ is the mean of K_S^k, and $\bar{m}$ the mean of $K = [K_S, K_T]$. $\mathbf{U}$ is obtained by replacing I in (33) with K.

(36) can be optimized in a similar way as (30).

F. Domain Transferability Estimation (DTE)

When there are a large number of source domains, estimating domain transferability can advise which domains are more important, and also reduce the computational cost. In BCIs, DTE can be used to find subjects which have low correlations to the tasks and hence may cause negative transfer. Although source domain selection is important, it is very challenging, and hence very few publications can be found in the literature [4], [14], [33], [34].

Next, we propose an unsupervised DTE strategy.

Assume there are z labeled sources domains $\mathbb{S}_i = \{X_S^{(i)}, \mathbf{y}_S^{(i)}\}_{i=1}^{z}$, where $X_S^{(i)}$ is the feature matrix of the i-th source domain, $\mathbf{y}_S^{(i)}$ is the corresponding label vector. Assume also there is a target domain $\mathbb{T}$ with unlabeled feature matrix X_T. Let S_b be the between-class scatter matrix, similar to S_b in (25), and $S_b^{\mathbb{S}_i,\mathbb{T}}$ be the scatter matrix between the source and the target domains. We define the discriminability of the i-th source domain as $DIS(\mathbb{S}_i) = \|S_b^{\mathbb{S}_i}\|_1$, and the difference between the source domain and the target domain as $DIF(\mathbb{S}_i, \mathbb{T}) = \|S_b^{\mathbb{S}_i,\mathbb{T}}\|_1$.

Then, the transferability of Source Domain $\mathbb{S}_i$ is computed as:

$$r(\mathbb{S}_i, \mathbb{T}) = \frac{DIS(\mathbb{S}_i)}{DIF(\mathbb{S}_i, \mathbb{T})} \tag{37}$$

We then select $z^* \in (1, z)$ source subjects with the highest $r(\mathbb{S}_i, \mathbb{T})$.

TABLE I
STATISTICS OF THE TWO MI AND TWO ERP DATASETS

Dataset	Number of Subjects	Number of Channels	Number of Time Samples	Trails per Subject	Class-Imbalance
MI1	7	59	300	200	No
MI2	9	22	750	144	No
RSVP	11	8	45	368-565	Yes
ERN	16	56	260	340	Yes

IV. EXPERIMENTS

In this section, we evaluate our method for both single-source to single-target (STS) transfers and multi-source to single-target (MTS) transfers. The code is available online.[2]

A. Datasets

We used two MI datasets and two ERP datasets in our experiments. Their statistics are summarized in Table I.

For both MI datasets, a subject sat in front of a computer screen. At the beginning of a trial, a fixation cross appeared on the black screen to prompt the subject to be prepared. Shortly after, an arrow pointing to a certain direction was presented as a visual cue for a few seconds, during which the subject performed a specific MI task. Then, the visual cue disappeared, and the next trial started after a short break. EEG signal was recorded during the experiment, and used to classify which MI the user was performing. Usually, EEG shortly after the visual cue onset is highly related to the MI task.

For the first MI dataset[3] (MI1), 59-channel EEGs were recorded at 100 Hz from seven healthy subjects, each with 100 left hand MIs and 100 right hand MIs. For the second MI dataset[4] (MI2), 22-channel EEGs were recorded at 250 Hz

[2]https://github.com/chamwen/MEKT
[3]http://www.bbci.de/competition/iv/desc_1.html.
[4]http://www.bbci.de/competition/iv/desc_2a.pdf.

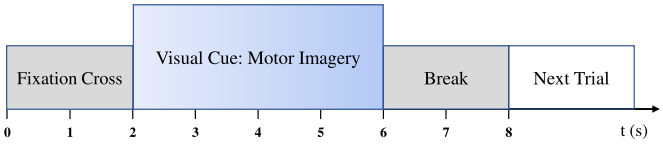

Fig. 2. Timing scheme of the motor imagery tasks in the first two datasets.

from nine heathy subjects, each with 72 left hand MIs and 72 right hand MIs. Both datasets were used for two-class classification.

The first ERP dataset[5] contained 8-channel EEG recordings from 11 healthy subjects in a rapid serial visual presentation (RSVP) experiment. The images were presented at different rates (5, 6, and 10 Hz) in three different experiments. We only used the 5 Hz version. The goal was to classify from EEG if the subject had seen a target image (with airplane) and non-target image (without airplane). The number of images for different subjects varying between 368 and 565, and the target to non-target ratio was around 1:9. The sampling rate was 2048 Hz, and the RSVP data had been band-pass filtered to 0.15-28 Hz.

The second ERP dataset[6] was recorded from a feedback error-related negativity (ERN) experiment [35], which was used in a Kaggle competition for two-class classification. It was collected from 26 subjects and partitioned into training set (16 subjects) and test set (10 subjects). We only used the 16 subjects in the training set as we do not have access to the test set. The average target to non-target ratio was around 1:4. The 56-channel EEG data had been downsampled to 200 Hz.

B. EEG Data Preprocessing

EEG signals from all datasets were preprocessed using the EEGLAB toolbox [36]. We followed the same preprocessing procedures in [24], [37].

For the two MI datasets, a causal 50-order 8-30 Hz.[7] finite impulse response (FIR) band-pass filter was used to remove muscle artifacts and direct current drift, and hence to obtain cleaner MI signals. Next, EEG signals between [0.5, 3.5] seconds after the cue onsets were extracted as trials. The RSVP signal was downsampled to 64 Hz to reduce the computational cost, and epoched to 0.7s intervals immediately after the stimulus onsets as trials. The ERN signal was bandpass filtered to 1-40 Hz, and epoched to 1.3s intervals immediately after the feedbacks (which contained the ERP associated with the user's response to the feedback event) as trials.

MI1 had 59 EEG channels, which were not easy to manipulate. Thus, we reduced the number of its tangent space features to the number of source domain samples (200), according to their F values in one-way ANOVA. For the ERN dataset, we used xDAWN [39] to reduce the number of channels from 56 to 6.

[5]https://www.physionet.org/physiobank/database/ltrsvp/.

[6]https://www.kaggle.com/c/inria-bci-challenge.

[7]We bandpass filtered the EEG signal to 8-30 Hz because MI is mainly indicated by the change of the mu rhythm (about 8–13 Hz) and the beta (about 14–30 Hz) rhythm [38]

TABLE II
INPUT SPACE DIMENSIONALITIES IN DIFFERENT STS TASKS

	MI1	MI2	ERP (RSVP and ERN)
Euclidean	6×200	6×144	$20\times n_i$
Tangent	200×200	253×144	$c^2 \times n_i$
Riemannian	$59\times59\times200$	$22\times22\times144$	$2c \times 2c \times n_i$

The dimensionalities of different input spaces are shown in Table II. n_i is the number of samples in the i-th domain, and c the number of selected channels for the two ERP datasets. Specifically, for RSVP, $c = 8$ and n_i varies from 368 to 565; for ERN, $c = 6$ and $n_i = 340$. Augmented covariance matrices [11] were used in the Riemannian space for ERP, so they had dimensionality of $2c \times 2c$. The $c \times c$ upper right block of the augmented covariance matrix contains temporal information [11], so these c^2 elements were selected as the tangent space features.

Next, we describe how the Euclidean space features were determined. For the two MI datasets, six log-variance features of the CSP filtered trials [see (4)] were used as features. For the two ERP datasets, after spatial filtering by xDAWN, we assembled each EEG trail (which is a matrix) into a vector, performed principal component analysis on all vectors from the source subjects, and extracted the scores for the first 20 principal components as features.

C. Baseline Algorithms

We compared our MEKT approaches (MEKT-R: the Riemannian mean is used as the reference matrix; MEKT-E: the Euclidean mean is used as the reference matrix; MEKT-L: the Log-Euclidean mean is used as the reference matrix) with seven state-of-the-art baseline algorithms for BCI classification. According to the feature space type, these baselines can be divided into three categories:

1) *Euclidean* space approaches:
   a) CSP-LDA (linear discriminant analysis) [40] for MI, and CSP-SVM (support vector machine) [41] for ERP.
   b) EA-CSP-LDA for MI, and EA-xDAWN-SVM for ERP, i.e., we performed EA [24] as a preprocessing step before spatial filtering and classification.
2) *Riemannian* space approach: RA-MDM [23] for MI, and xDAWN-RA-MDM for ERP.
3) *Tangent* space approaches, which were proposed for computer vision applications, and have not been used in BCIs before. CA was used before each of them. In each learned subspace, the sLDA classifier [42] was used for MI, and SVM for ERP.
   a) CA (centroid alignment).
   b) CA-CORAL (correlation alignment) [13].
   c) CA-GFK (geodesic flow kernel) [14].
   d) CA-JDA (joint distribution adaptation) [15].
   e) CA-JGSA (joint geometrical and statistical alignment) [16].

Hyper-parameters of all baselines were set according to the recommendations in their corresponding publications. For

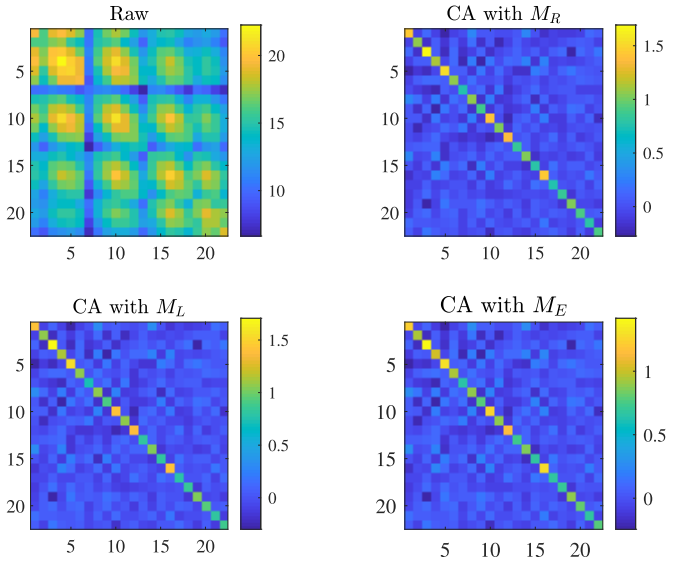

Fig. 3. The raw covariance matrix (Trial 1, Subject 1, MI2), and those after CA using different reference matrices.

MEKT, $T = 5$, $\alpha = 0.01$, $\beta = 0.1$, $\rho = 20$, and $d = 10$ were used.

D. Experimental Settings

We evaluated unsupervised STS and MTS transfers. In STS, one subject was selected as the target, and another as the source. Let z be the number of subjects in a dataset. Then, there were $z(z-1)$ different STS tasks. In MTS, one subject was used as the target, and all others as the sources, so there were z different MTS tasks. For example, MI1 included seven subjects, so we had $7 \times 6 = 42$ STS tasks, e.g., S2→S1 (Subject 2 as the source, and Subject 1 as the target), S3→S1, S4→S1, S5→S1, S6→S1, S7→S1, …S6→S7, and seven MTS tasks, e.g., {S2, S3, S4, S5, S6, S7}→S1, …, {S1, S2, S3, S4, S5, S6}→S7.

The balanced classification accuracy (BCA) was used as the performance measure:

$$BCA = \frac{1}{l}\sum_{k=1}^{l}\frac{tp_k}{n_k}, \tag{38}$$

where tp_k and n_k are the number of true positives and the number of samples in Class k, respectively.

E. Visualization

As explained in Section III-A, CA makes the aligned covariance matrices approximate the identity matrix, no matter whether the Riemannian mean, or the Euclidean mean, or the Log-Euclidean mean, is used as the reference matrix. To demonstrate that, Fig. 3 shows the raw covariance matrix of the first EEG trial of Subject 1 in MI2, and the aligned covariance matrices using different references. The raw covariance matrix is nowhere close to identity, but after CA, the covariance matrices are approximately identity, and hence the corresponding EEG trials are approximately whitened.

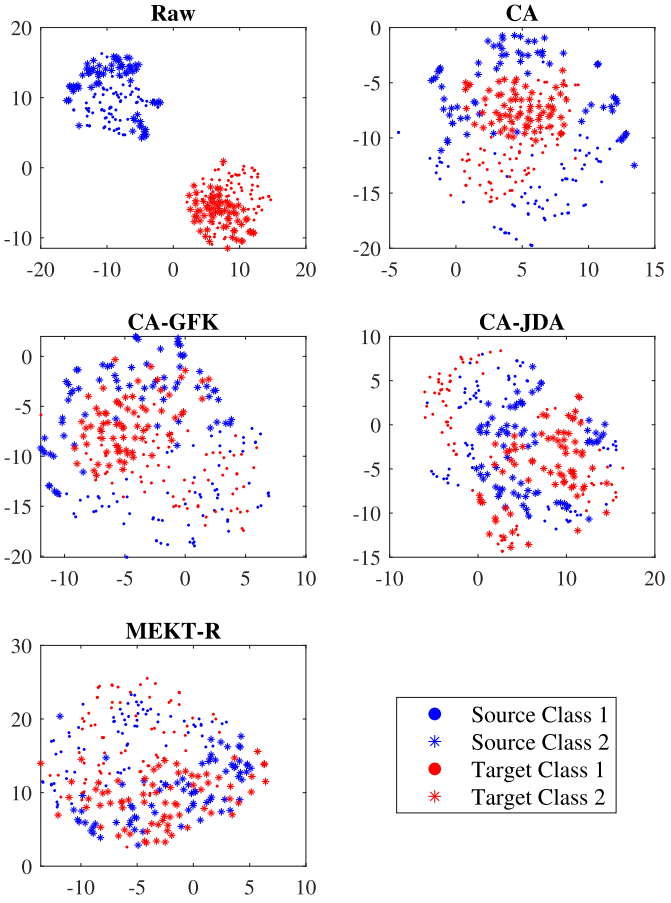

Fig. 4. *t*-SNE visualization of the data distributions before and after CA, and with different transfer learning approaches, when transferring Subject 2's data (source) to Subject 1 (target) in MI2.

Next, we used t-SNE [43] to reduce the dimensionality of the EEG trials to two, and visualize if MEKT can bring the data distributions of the source and the target domains together. Fig. 4 shows the results on transferring Subject 2's data to Subject 1 in MI2, before and after different data alignment approaches. Before CA, the source domain and target domain samples do not overlap at all. After CA, the two sets of samples have identical mean, but different variances. CA-GFK and CA-JDA make the variance of the source domain samples and the variance of the target domain samples approximately identical, but different classes are still not well separated. MEKT-R not only makes the overall distributions of the source domain samples and the target domain samples consistent, but also samples from the same class in the two domains close, which should benefit the classification.

F. Classification Accuracies

The means and standard deviations of the BCAs on the four datasets with STS and MTS transfers are shown in Tables III and IV, respectively. All MEKT-based approaches achieved the best (in bold) or the second best (underlined) performance in all scenarios in contrast to the baselines.

Fig. 5 shows the BCAs of all tangent space based approaches when different reference matrices were used in

TABLE III
MEAN (%) AND STANDARD DEVIATION (%; IN PARENTHESIS) OF THE BCAS IN STS TRANSFERS. FOR THE CA-BASED APPROACHES, THE SLDA CLASSIFIER WAS USED FOR MI, AND SVM FOR ERP

	MI1	MI2	Avg
CSP-LDA	57.23 (10.56)	58.7 (11.58)	57.97
EA-CSP-LDA	66.85 (10.56)	65.00 (14.06)	65.93
RA-MDM	64.98 (10.37)	66.60 (12.60)	65.79
CA	66.17 (9.93)	66.02 (13.14)	66.10
CA-CORAL	67.69 (10.68)	67.26 (13.34)	67.48
CA-GFK	66.62 (10.53)	65.54 (13.56)	66.08
CA-JDA	66.01 (12.55)	66.59 (15.28)	66.30
CA-JGSA	65.81 (13.06)	65.90 (16.73)	65.85
MEKT-E	69.19 (12.84)	68.34 (15.51)	68.77
MEKT-L	70.74 (12.28)	68.56 (15.66)	69.65
MEKT-R	**70.99** (12.46)	**68.73** (15.73)	**69.86**
	RSVP	ERN	Avg
CSP-LDA	58.58 (7.98)	54.34 (5.87)	56.46
EA-CSP-LDA	58.76 (7.51)	55.57 (6.26)	57.17
RA-MDM	60.37 (8.05)	56.22 (6.89)	58.30
CA	58.34 (6.98)	56.97 (7.06)	57.66
CA-CORAL	58.45 (6.84)	57.04 (7.00)	57.75
CA-GFK	59.93 (7.61)	57.24 (7.34)	58.59
CA-JDA	60.27 (7.75)	57.56 (7.63)	58.92
CA-JGSA	55.23 (6.74)	57.17 (7.72)	56.20
MEKT-E	61.08 (8.59)	**58.01** (7.76)	**59.55**
MEKT-L	61.15 (8.44)	57.91 (7.74)	59.53
MEKT-R	**61.24** (8.36)	57.85 (7.75)	**59.55**

TABLE IV
MEAN (%) AND STANDARD DEVIATION (%; IN PARENTHESIS) OF THE BCAS IN MTS TRANSFERS

	MI1	MI2	Avg
CSP-LDA	59.71 (12.93)	67.75 (12.92)	63.73
EA-CSP-LDA	79.79 (6.57)	73.53 (15.96)	76.66
RA-MDM	73.29 (9.25)	72.07 (9.88)	72.68
CA	76.29 (9.66)	71.84 (13.89)	74.07
CA-CORAL	78.86 (8.73)	72.38 (13.38)	75.62
CA-GFK	76.79 (12.57)	72.99 (15.82)	74.89
CA-JDA	81.07 (11.19)	74.15 (15.77)	77.61
CA-JGSA	76.79 (12.35)	73.07 (16.33)	74.93
MEKT-E	81.29 (10.18)	76.00 (17.61)	78.65
MEKT-L	83.07 (9.30)	**76.54** (16.72)	79.81
MEKT-R	**83.42** (9.55)	76.31 (16.76)	**79.87**
	RSVP	ERN	Avg
CSP-LDA	65.36 (9.32)	61.87 (4.51)	63.62
EA-CSP-LDA	**69.07** (9.05)	64.63 (5.86)	66.85
RA-MDM	67.29 (8.38)	62.90 (6.79)	65.10
CA	67.35 (7.52)	65.89 (7.30)	66.62
CA-CORAL	66.94 (7.46)	66.17 (7.74)	66.56
CA-GFK	67.75 (7.48)	66.03 (7.50)	66.89
CA-JDA	66.06 (6.18)	64.64 (6.50)	65.35
CA-JGSA	64.57 (5.79)	57.68 (8.04)	61.13
MEKT-E	67.92 (6.70)	**66.70** (8.00)	**67.31**
MEKT-L	68.40 (6.40)	65.98 (7.94)	67.19
MEKT-R	68.38 (6.36)	66.17 (7.68)	67.28

CA. The Riemannian mean obtained the best BCA in four out of the six approaches, and also the best overall performance.

We also performed paired t-tests on the BCAs to check if the performance improvements of MEKT-R over others were statistically significant. Before each t-test, we performed a Lilliefors test [44] to verify that the null hypothesis that the data come from a normal distribution cannot be rejected. Then, we performed false discovery rate corrections [45] by

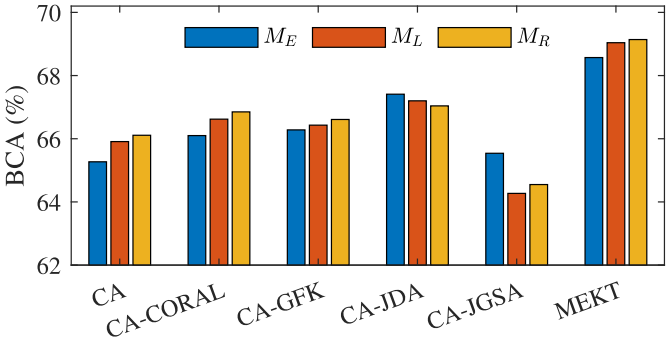

Fig. 5. Average BCAs (%) of the tangent space approaches on the four datasets, when different reference matrices were used in CA.

TABLE V
FALSE DISCOVERY RATE ADJUSTED p-VALUES IN PAIRED t-TESTS ($\alpha = 0.05$)

	MEKT-R vs	MI1	MI2	RSVP	ERN
STS	CSP-LDA	**.0000**	**.0000**	–	–
	xDAWN-SVM	–	–	**.0002**	**.0000**
	EA-CSP-LDA	**.0030**	**.0003**	–	–
	EA-xDAWN-SVM	–	–	**.0000**	**.0000**
	RA-MDM	**.0003**	**.0340**	**.0412**	**.0004**
	CA	**.0000**	**.0006**	**.0000**	**.0010**
	CA-CORAL	**.0005**	**.0340**	**.0000**	**.0014**
	CA-GFK	**.0000**	**.0001**	**.0016**	**.0130**
	CA-JDA	**.0003**	**.0183**	**.0386**	.2627
	CA-JGSA	**.0021**	**.0006**	**.0000**	**.0241**
MTS	CSP-LDA	**.0329**	.1140	–	–
	xDAWN-SVM	–	–	.2077	**.0306**
	EA-CSP-LDA	.2808	.1636	–	–
	EA-xDAWN-SVM	–	–	.5733	.2632
	xDAWN-RA-MDM	.0824	.1636	.5347	.0632
	CA	**.0329**	.1260	.4727	.8380
	CA-CORAL	.0897	.1636	.3477	.9914
	CA-GFK	.0824	.1260	.5347	.9117
	CA-JDA	.2379	.1636	**.0349**	.0632
	CA-JGSA	.1344	.1636	**.0323**	**.0018**

a linear-step up procedure under a fixed significance level ($\alpha = 0.05$) on the paired p-values of each task.

The false discovery rate adjusted p-values (q-values) are shown in Table V. MEKT-R significantly outperformed all baselines in almost all STS transfers. The performance improvements became less significant when there were multiple source domains, which is reasonable, because generally in machine learning the differences between different algorithms diminish as the amount of training data increases.

We also considered linear and radial basis function (RBF; kernel width 0.1) kernels in MEKT-R, and repeated the above experiments. The results are shown in Table VI, where *Primal* denotes the primal MEKT-R without kernelization. The primal MEKT-R achieved the best (in bold) or the second best (underlined) performance in all scenarios. However, the differences among the three approaches were very small.

G. Computational Cost

This subsection empirically checked the computational cost of different algorithms, which were implemented in Matlab 2018a on a laptop with i7-8550U CPU@2.00GHz, 8GB memory, running 64-bit Windows 10 Education Edition.

TABLE VI
AVERAGE BCAS(%) OF THE PROPOSED MEKT UNDER DIFFERENT KERNELS

		Primal	Linear	RBF
STS	MI1	**70.99**	**70.99**	70.37
	MI2	**68.73**	**68.73**	68.37
	RSVP	61.24	60.49	**61.78**
	ERN	57.85	57.44	**58.45**
MTS	MI1	**83.42**	83.36	78.21
	MI2	**76.31**	**76.31**	76.08
	RSVP	68.38	**68.41**	68.22
	ERN	**66.17**	66.02	65.49
	Avg	**69.14**	68.97	68.37

TABLE VII
COMPUTING TIME (SECONDS) OF DIFFERENT APPROACHES IN STS AND MTS TRANSFERS

		RA-MDM	EA	CA-JDA	MEKT-E	MEKT-L	MEKT-R
STS	MI1	5.49	**0.44**	5.45	2.53	2.75	5.42
	MI2	0.48	**0.27**	0.54	0.43	0.47	0.60
	RSVP	0.42	**0.05**	0.45	0.23	0.27	0.43
	ERN	0.54	0.47	0.53	**0.38**	0.42	0.53
MTS	MI1	13.61	**0.94**	9.24	11.06	11.48	12.96
	MI2	1.01	**0.69**	1.35	1.13	1.20	1.29
	RSVP	3.13	**1.08**	8.64	5.61	5.98	6.74
	ERN	**5.49**	7.95	14.92	10.39	10.74	11.95

For simplicity, we only selected one transfer task in each dataset. For STS transfer, the first subject in each dataset was selected as the target domain, and the second subject as the source domain. For MTS transfer, the first subject as the target domain, and all other subjects as the source domains. we repeated the experiment 20 times, and show the average computing time in Table VII. In summary, EA was the most efficient. RA-MDM, CA-JDA and MEKT-R had similar computational cost. MEKT-L and MEKT-E had comparable classification performance with MEKT-R (Tables III and IV), but much less computational cost. MEKT-L achieved the best compromise between the classification accuracy and the computational cost.

H. Effectiveness of the Joint Probability MMD

To validate the superiority of the joint probability MMD over the traditional MMD, we replaced the joint probability MMD term $\mathcal{D}'_{S,T}$ in (30) by the traditional MMD term $\mathcal{D}_{S,T}$ in (21), and repeated the experiments. The results are shown in Table VIII. The joint probability MMD outperformed the traditional MMD in six out of the eight tasks. We expect that the joint probability MMD should also be advantageous in other applications that the traditional MMD is now used.

I. Effectiveness of DTE

This subsection validates our DTE strategy on MTS tasks to select the most beneficial source subjects.

Table IX shows the BCAs when different source domain selection approaches were used: RAND randomly selected $round[(z-1)/2]$ source subjects [because there was randomness, we repeated the experiment 20 times, and report the mean and standard deviation (in the parentheses)], ROD was the approach proposed in [14], and ALL used all z source subjects. Table X shows the computational cost of different algorithms.

Tables IX and X shows that the proposed DTE outperformed RAND and ROD in terms of the classification accuracy. Although its BCAs were generally slightly worse than those of ALL, its computational cost was much lower than ALL, especially when z became large, i.e., when $z \gg 1$, it can save over 50% computational cost.

TABLE VIII
AVERAGE BCAS (%) WHEN $\mathcal{D}_{S,T}$ IN (21) OR $\mathcal{D}'_{S,T}$ IN (22) WAS USED IN (30)

		$\mathcal{D}_{S,T}$	$\mathcal{D}'_{S,T}$
STS	MI1	65.33	**70.99**
	MI2	66.78	**68.73**
	RSVP	61.11	**61.24**
	ERN	**58.62**	57.85
MTS	MI1	73.86	**83.42**
	MI2	74.23	**76.31**
	RSVP	**69.33**	68.38
	ERN	65.59	**66.17**
	Avg	66.86	**69.14**

TABLE IX
AVERAGE BCAS (%) WITH DIFFERENT SOURCE DOMAIN SELECTION APPROACHES. RAND, ROD AND DTE EACH SELECTED $round[(z-1)/2]$ SOURCE SUBJECTS. ALL USED ALL SOURCE SUBJECTS

	z	RAND	ROD	DTE	ALL
MI1	7	81.53 (1.19)	81.86	82.14	**83.42**
MI2	9	75.05 (1.06)	74.38	76.23	**76.31**
RSVP	11	67.48 (0.31)	67.79	**68.70**	68.38
ERN	16	65.31 (0.52)	65.36	65.51	**66.17**

TABLE X
COMPUTING TIME (SECONDS) OF DIFFERENT SOURCE DOMAIN SELECTION APPROACHES. RAND, ROD AND DTE EACH SELECTED $round[(z-1)/2]$ SOURCE SUBJECTS. ALL USED ALL SOURCE SUBJECTS

	z	RAND	ROD	DTE	ALL
MI1	7	**11.55**	12.46	11.77	12.84
MI2	9	**0.90**	1.11	0.94	1.24
RSVP	11	**3.08**	3.22	3.10	6.80
ERN	16	**6.27**	6.42	6.29	11.57

V. CONCLUSIONS

Transfer learning is popular in EEG-based BCIs to cope with variations among different subjects and/or tasks. This paper has considered offline unsupervised cross-subject EEG classification, i.e., we have labeled EEG trials from one or more source subjects, but only unlabeled EEG trials from the target subject. We proposed a novel MEKT approach, which has three steps: 1) align the covariance matrices of the EEG trials in the Riemannian manifold; 2) extract tangent space features; and, 3) perform domain adaptation by minimizing the joint probability distribution shift between the source and the target domains, while preserving their geometric structures. An optional fourth step, DTE, was also proposed to identify the most beneficial source domains, and hence

to reduce the computational cost. Experiments on four EEG datasets from two different BCI paradigms demonstrated that MEKT outperformed several state-of-the-art transfer learning approaches. Moreover, DTE can reduce more than half of the computational cost when the number of source subjects is large, with little sacrifice of classification accuracy.

REFERENCES

[1] J. Wolpaw, N. Birbaumer, D. McFarland, G. Pfurtscheller, and T. Vaughan, "Brain-computer interfaces for communication and control," *Clin. Neurophys.*, vol. 113, no. 6, pp. 767–791, 2002.

[2] R. P. Rao, *Brain-Computer Interfacing: Introduction*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[3] B. He, B. Baxter, B. J. Edelman, C. C. Cline, and W. W. Ye, "Noninvasive brain-computer interfaces based on sensorimotor rhythms," *Proc. IEEE*, vol. 103, no. 6, pp. 907–925, Jun. 2015.

[4] D. Wu, "Online and offline domain adaptation for reducing BCI calibration effort," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 4, pp. 550–563, Aug. 2017.

[5] F. Lotte *et al.*, "A review of classification algorithms for EEG-based brain–computer interfaces: A 10 year update," *J. Neural Eng.*, vol. 15, no. 3, Jun. 2018, Art. no. 031005.

[6] Z. J. Koles, M. S. Lazar, and S. Z. Zhou, "Spatial patterns underlying population differences in the background EEG," *Brain Topography*, vol. 2, no. 4, pp. 275–284, 1990.

[7] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache, "Fast and simple calculus on tensors in the log-Euclidean framework," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent*, Palm Springs, CA, USA, Oct. 2005, pp. 115–122.

[8] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Multiclass Brain–Computer interface classification by Riemannian geometry," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 4, pp. 920–928, Apr. 2012.

[9] F. Yger, M. Berar, and F. Lotte, "Riemannian approaches in brain-computer interfaces: A review," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1753–1762, Oct. 2017.

[10] P. L. C. Rodrigues, C. Jutten, and M. Congedo, "Riemannian procrustes analysis: Transfer learning for Brain–Computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 66, no. 8, pp. 2390–2401, Aug. 2019.

[11] L. Korczowski, M. Congedo, and C. Jutten, "Single-trial classification of multi-user P300-based brain-computer interface using Riemannian geometry," in *Proc. 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Milan, Italy, Aug. 2015, pp. 1769–1772.

[12] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Netw.*, vol. 22, no. 2, pp. 199–210, Feb. 2011.

[13] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. 30th AAAI Conf. Artif. Intell.*, Arizona, USA, Mar. 2016, pp. 2058–2065.

[14] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic flow kernel for unsupervised domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 2066–2073.

[15] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu, "Transfer feature learning with joint distribution adaptation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, Australia, Dec. 2013, pp. 2200–2207.

[16] J. Zhang, W. Li, and P. Ogunbona, "Joint geometrical and statistical alignment for visual domain adaptation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1859–1867.

[17] D. Wu, B. J. Lance, and T. D. Parsons, "Collaborative filtering for brain-computer interaction using transfer learning and active class selection," *PLoS ONE*, vol. 8, no. 2, 2013, Art. no. e56624.

[18] D. Wu, V. J. Lawhern, W. D. Hairston, and B. J. Lance, "Switching EEG headsets made easy: Reducing offline calibration effort using active weighted adaptation regularization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 24, no. 11, pp. 1125–1137, Nov. 2016.

[19] V. Jayaram, M. Alamgir, Y. Altun, B. Scholkopf, and M. Grosse-Wentrup, "Transfer learning in brain-computer interfaces," *IEEE Comput. Intell. Mag.*, vol. 11, no. 1, pp. 20–31, Feb. 2016.

[20] H. Kang, Y. Nam, and S. Choi, "Composite common spatial pattern for Subject-to-Subject transfer," *IEEE Signal Process. Lett.*, vol. 16, no. 8, pp. 683–686, Aug. 2009.

[21] F. Lotte and C. Guan, "Learning from other subjects helps reducing brain-computer interface calibration time," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 614–617.

[22] Y. Jin, M. Mousavi, and V. R. de Sa, "Adaptive CSP with subspace alignment for subject-to-subject transfer in motor imagery brain-computer interfaces," in *Proc. 6th Int. Conf. Brain-Comput. Interface (BCI)*, GangWon, South Korea, Jan. 2018, pp. 1–4.

[23] P. Zanini, M. Congedo, C. Jutten, S. Said, and Y. Berthoumieu, "Transfer learning: A Riemannian geometry framework with applications to Brain–Computer interfaces," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 5, pp. 1107–1116, May 2018.

[24] H. He and D. Wu, "Transfer learning for Brain–Computer interfaces: A Euclidean space data alignment approach," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 399–410, Feb. 2020.

[25] A. Barachant, S. Bonnet, M. Congedo, and C. Jutten, "Classification of covariance matrices using a riemannian-based kernel for BCI applications," *Neurocomputing*, vol. 112, pp. 172–178, Jul. 2013.

[26] H. Ramoser, J. Müller-Gerking, and G. Pfurtscheller, "Optimal spatial filtering of single trial EEG during imagined hand movement," *IEEE Trans. Rehabil. Eng.*, vol. 8, no. 4, pp. 441–446, Dec. 2000.

[27] R. Bhatia, *Positive Definite Matrices*. Princeton, NJ, USA: Princeton Univ. Press, 2009.

[28] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, Mar. 2012.

[29] M. Belkin and P. Niyogi, "Semi-supervised learning on Riemannian manifolds," *Mach. Learn.*, vol. 56, nos. 1–3, pp. 209–239, Jul. 2004.

[30] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 12, pp. 1624–1637, Dec. 2005.

[31] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[32] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, Nov. 2006.

[33] D. Wu, V. J. Lawhern, S. Gordon, B. J. Lance, and C.-T. Lin, "Driver drowsiness estimation from EEG signals using online weighted adaptation regularization for regression (OwARR)," *IEEE Trans. Fuzzy Syst.*, vol. 25, no. 6, pp. 1522–1535, Dec. 2017.

[34] C.-S. Wei, Y.-P. Lin, Y.-T. Wang, T.-P. Jung, N. Bigdely-Shamlo, and C.-T. Lin, "Selective transfer learning for EEG-based drowsiness detection," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Hong Kong, Oct. 2015, pp. 3229–3232.

[35] P. Margaux, M. Emmanuel, D. Sbastien, B. Olivier, and M. Jrmie, "Objective and subjective evaluation of online error correction during P300-based spelling," *Adv. Hum.-Comput. Interact.*, vol. 2012, Dec. 2012, Art. no. 578295.

[36] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[37] X. Zhang and D. Wu, "On the vulnerability of CNN classifiers in EEG-based BCIs," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 5, pp. 814–825, May 2019.

[38] Q. Ai, Q. Liu, W. Meng, and S. Q. Xie, *Advanced Rehabilitative Technology*. New York, NY, USA: Academic, 2018.

[39] B. Rivet, A. Souloumiac, V. Attina, and G. Gibert, "XDAWN algorithm to enhance evoked potentials: Application to Brain–Computer interface," *IEEE Trans. Biomed. Eng.*, vol. 56, no. 8, pp. 2035–2043, Aug. 2009.

[40] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.

[41] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, Apr. 2011.

[42] R. Peck and J. Van Ness, "The use of shrinkage estimators in linear discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vols. PAMI–4, no. 5, pp. 530–537, Sep. 1982.

[43] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

[44] H. W. Lilliefors, "On the kolmogorov-smirnov test for normality with mean and variance unknown," *J. Amer. Stat. Assoc.*, vol. 62, no. 318, pp. 399–402, Jun. 1967.

[45] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: A practical and powerful approach to multiple testing," *J. Roy. Stat. Soc., Ser. B, (Methodol.)*, vol. 57, no. 1, pp. 289–300, Jan. 1995.