Aligning Infinite-Dimensional Covariance Matrices in Reproducing Kernel Hilbert Spaces for Domain Adaption: Supplementary Material

Zhen Zhang, Mianzhi Wang, Yan Huang, Arye Nehorai
Washington University in St. Louis
{zhen.zhang, mianzhi.wang, yanhuang640, nehorai}@wustl.edu

Abstract

The supplementary material consists of three parts. In the first part, we provide insights on our framework by discussing the expressiveness of covariance descriptors in RKHS. In the second part, we provide more discussion on the experiments. In the third part, we prove all the mathematical results presented in the paper.

1. Discussion on RKHS covariance descriptors

Given RKHS data matrix $\Phi_X = [\phi(x_1), \phi(x_2), ..., \phi(x_N)]$, the maximum likelihood estimation of the RKHS covariance descriptor is

$$\mathrm{M}C = \frac{1}{N}\sum_{i=1}^{N}\left[\phi(x_i) - \hat{\mu}\right]\left[\phi(x_i) - \hat{\mu}\right]^T = \Phi_{\boldsymbol{X}}\boldsymbol{J}_N\boldsymbol{J}_N^T\Phi_{\boldsymbol{X}}^T, \tag{1}$$

where $\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N}\phi(x_i)$ is the empirical mean of samples. If we use the linear kernel, *i.e.*, $k(\vec{\boldsymbol{x}}, \vec{\boldsymbol{y}}) = \vec{\boldsymbol{x}}^T\vec{\boldsymbol{y}}$, then the corresponding feature map is just the identity function, *i.e.*, $\phi(\vec{\boldsymbol{x}}) = \vec{\boldsymbol{x}}$. As a result, the expression (1) degenerates to the MLE of covariance matrices in $\mathbb{R}^n$. For the computationally efficient estimation (See Section 3.2 in the paper), we have the same explanation. Therefore, RKHS covariance descriptors are the generalization of covariance matrices.

From the standpoint of kernel methods, with the nonlinear kernel, *e.g.*, the RBF kernel, RKHS covariance descriptors, which can capture nonlinear structure and higher-order correlations, are more informative than the covariance matrices. When aligning two infinite-dimensional covariance descriptors, we in fact match infinitely many orders of statistics.

Similar strategy of employing RKHS covariance descriptors to characterize a set of samples can be found in [1, 3], where the authors represent each image by a covariance descriptor in RKHS and quantify the discrepancies between covariance descriptors to classify images.

2. Discussion on the experiments

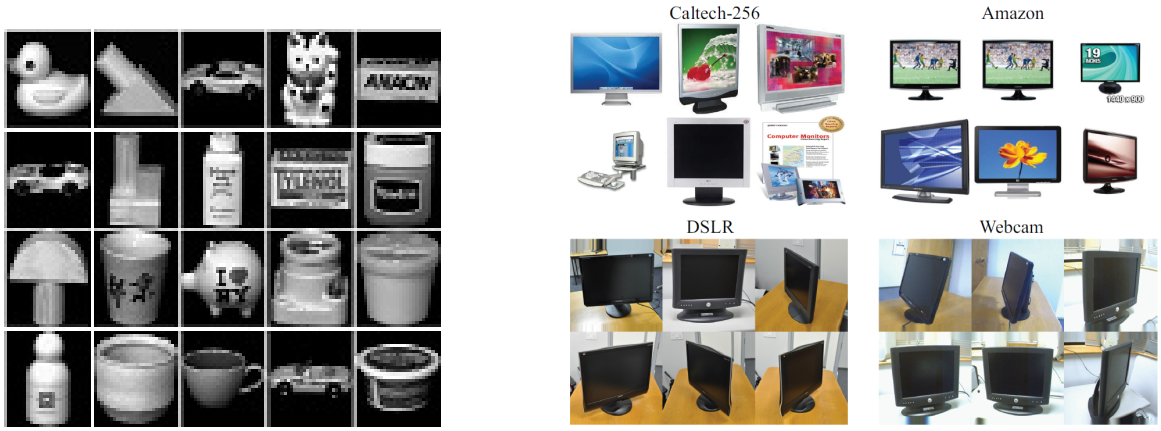
2.1. More about the datasets

Fig. 1(a) shows sample images in the COIL20 dataset. For each object in COIL20, we provide one example image of the 72 total. Fig. 1(b) shows sample images from the monitor category in the Office-Caltech dataset.

Table 1 lists the top categories and subcategories in the 20-Newsgroups dataset.

2.2. Visualization using kernel PCA

We use kernel principal components analysis [4] (kernel PCA) to visualize source and target samples in RKHS. We implement experiments on a cross-domain dataset generated from 20-Newsgroups. The source dataset consists of four subcategories of Comp and Rec: comp.graphics, comp.sys.mac.hardware, rec.sport.baseball, and rec.sport.hokey. The target dataset consists of the other four subcategories: comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, rec.autos, and rec.motorcycles. So there are $970+958+991+997 = 3916$ samples in the source domain, and $963+979+987+993 = 3922$



(a) (b)

Figure 1: (a) Sample images from the COIL20 dataset. (b) Sample images from the Office-Caltech dataset.

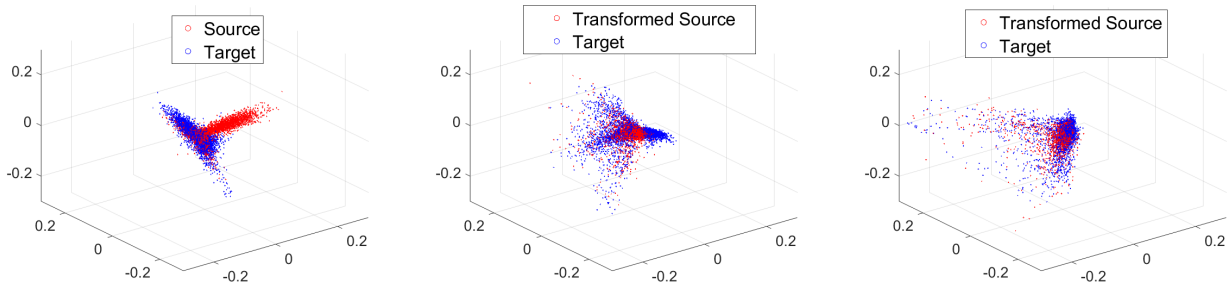
Table 1: Top categories and subcategories in the 20-Newsgroups dataset. All features have a dimensional of 25,804.

Top Category	Subcategory	Number of Samples
Comp	comp.graphics	970
	comp.os.ms-windows.misc	963
	comp.sys.ibm.pc.hardware	979
	comp.sys.mac.hardware	958
Rec	rec.autos	987
	rec.motorcycles	993
	rec.sport.baseball	991
	rec.sport.hokey	997
Sci	sci.crypt	989
	sci.electronics	984
	sci.med	987
	sci.space	985
Talk	talk.politics.mideast	940
	talk.politics.misc	774
	talk.politics.guns	909
	talk.region.misc	627

samples in the target domain. We employ the RBF kernel, and visualize the coefficients of RKHS samples with respect to the first three principal components. Fig. 2(a) shows the results with the non-adapted kernel matrix $\boldsymbol{K}$. We can see that the source and target distributions are very different. Fig. 2(b) and Fig. 2(c) show the results with the domain-invariant kernel matrices $\mathrm{WC}\tilde{\boldsymbol{K}}$ and $\mathrm{OT}\tilde{\boldsymbol{K}}$ (see Section 5 in the paper), respectively. From Fig. 2(b) and (c), we conclude that after "moving" the source data by the kernel whitening-coloring map or the kernel optimal transport map, the transformed source samples and target samples are closely distributed in RKHS. In addition, we note that the recognition accuracies with non-adapted kernel matrix $\boldsymbol{K}$ and domain-invariant kernel matrices $\mathrm{WC}\tilde{\boldsymbol{K}}$ and $\mathrm{OT}\tilde{\boldsymbol{K}}$ are ***61.19%***, ***94.11%*** and ***95.99%***, respectively. The highly superior performances of our approaches demonstrate the effectiveness of aligning covariance descriptors.

2.3. Out-of-sample generalization on the Reuters-21578 dataset

In this subsection, we measure our approaches' ability to generalize out-of-sample patterns. We follow the experimental protocol in [2], and conduct experiments on the preprocessed Reuters-21578 datasets. To train the model, we randomly select 500 labeled samples from the source domain and 300 unlabeled samples from the target domain. We test the model on the remaining unlabeled samples in the target domain. We repeat the above procedures 500 times, and report the average



(a) (b) (c)

Figure 2: We implement experiments on a cross-domain dataset generated from 20-Newsgroups. We use kernel PCA [4] to visualize the data in RKHS. (a) Visualization of the source and target samples with the non-adapted kernel. (b) Visualization of the transformed source and target samples with the domain-invariant kernel $\mathrm{WC}\tilde{\boldsymbol{K}}$. (c) Visualization of the transformed source and target samples with the domain-invariant kernel $\mathrm{OT}\tilde{\boldsymbol{K}}$.

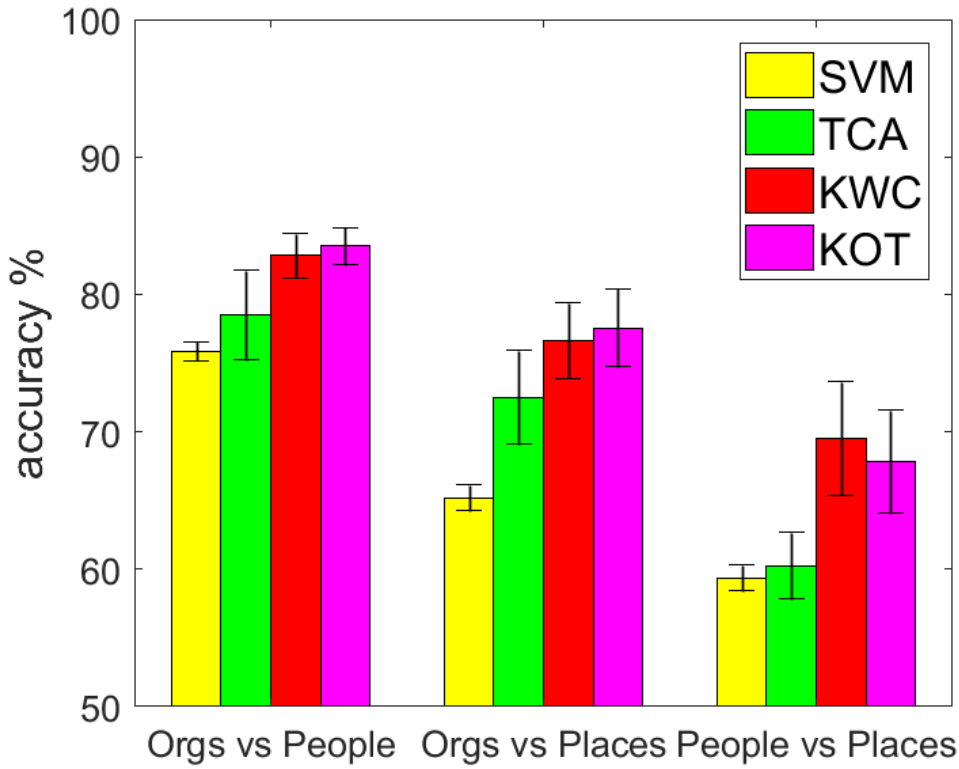


Figure 3: Accuracies and confidence intervals in recognizing out-of-sample data of the Reuters-21578 dataset. For all four methods, we use the linear kernel.

accuracies and standard errors. In these experiments, we compare our approaches with only the standard SVM and TCA, both of which possess generalizability. The parameters are set to be the same as those in the transductive setting. In Fig. 3 and Fig. 4, the experimental results with the both linear kernel and RBF kernel show that our approaches KWC and KOT outperform TCA and SVM with statistical significance.

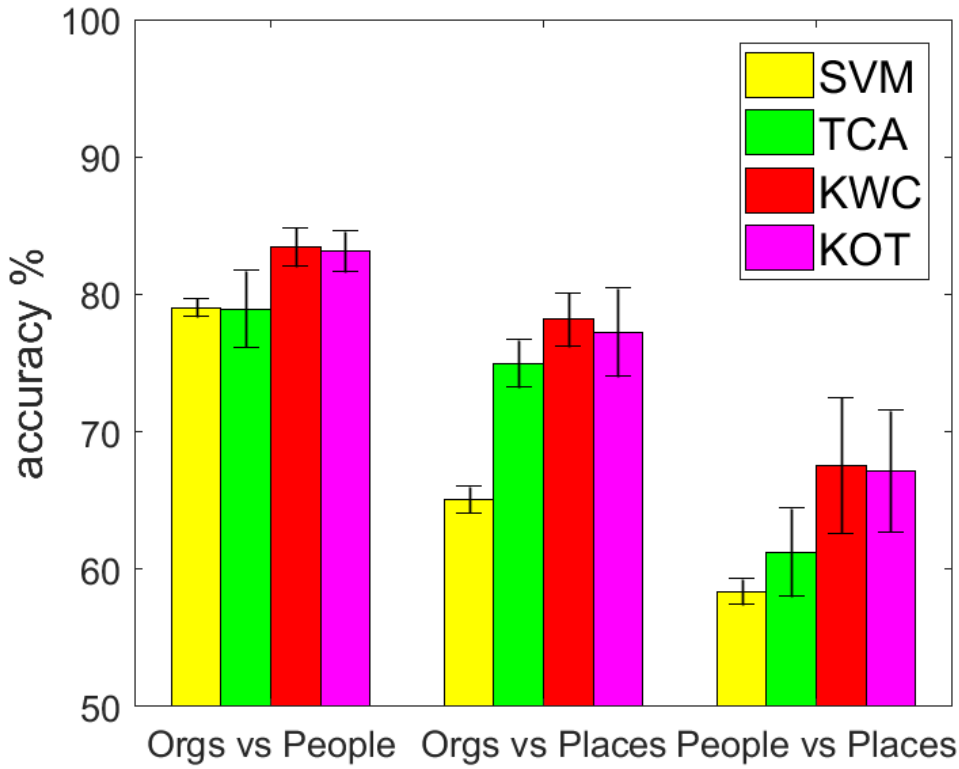


Figure 4: Accuracies and confidence intervals in recognizing out-of-sample data of the Reuters-21578 dataset. For all four methods, we use the RBF kernel.

3. Proofs of the mathematical results in the paper

We first provide some useful lemmas (corollaries), which will be frequently used.

Lemma 1. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two separable Hilbert spaces. Let $G : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator with finite rank, and let $G^* : \mathcal{H}_2 \to \mathcal{H}_1$ be its adjoint operator. Then* $\mathrm{Im}(G) = \mathrm{Im}(GG^*)$.

Proof. We need to show that $\mathrm{Im}(G) \subseteq \mathrm{Im}(GG^*)$, and $\mathrm{Im}(GG^*) \subseteq \mathrm{Im}(G)$.

1. Since $\mathrm{Im}(G) = \mathrm{Ker}^{\perp}(G^*)$ and $\mathrm{Im}(GG^*) = \mathrm{Ker}^{\perp}[(GG^*)^*] = \mathrm{Ker}^{\perp}(GG^*)$ [1], to obtain $\mathrm{Im}(G) \subseteq \mathrm{Im}(GG^*)$, it is sufficient to show that $\mathrm{Ker}(GG^*) \subseteq \mathrm{Ker}(G^*)$. $\forall v \in \mathrm{Ker}(GG^*)$, $GG^*v = 0 \implies \langle GG^*v, v\rangle_{\mathcal{H}_2} = \langle G^*v, G^*v\rangle_{\mathcal{H}_1} = 0 \implies G^*v = 0 \implies v \in \mathrm{Ker}(G^*)$. So $\mathrm{Ker}(GG^*) \subseteq \mathrm{Ker}(G^*)$.

2. $\forall v \in \mathrm{Im}(GG^*)$, there exists $w \in \mathcal{H}_2$, such that $v = GG^*w$. Writing $v = G(G^*w)$, we have $v \in \mathrm{Im}(G)$. So $\mathrm{Im}(GG^*) \subseteq \mathrm{Im}(G)$.

□

Lemma 2. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two separable Hilbert spaces. Let $G : \mathcal{H}_1 \to \mathcal{H}_2$ be a linear operator with finite rank, and let $G^* : \mathcal{H}_2 \to \mathcal{H}_1$ be its adjoint operator. Let $\{\lambda_k(G^*G)\}_{k=1}^r$ and $\{\varphi_k(G^*G)\}_{k=1}^r$ be the positive eigenvalues and the corresponding orthonormal eigenvectors of the operator G^*G. Then,*

$$GG^* = \sum_{k=1}^{r} \lambda_k(G^*G) \frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}} \otimes \frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}} \tag{2}$$

[1] Given two Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, and a linear operator $G : \mathcal{H}_1 \to \mathcal{H}_2$, the kernel space of G is defined as $\mathrm{Ker}(G) = \{v \in \mathcal{H}_1, Gv = 0_{\mathcal{H}_2}\}$.

is the orthogonal eigen-decomposition of the operator GG^*.

Remark 1. *The tensor product $\otimes$ is defined such that $(u \otimes v)w = u\langle v, w\rangle_{\mathcal{H}}, \forall u, v, w \in \mathcal{H}$, which is the analogy of the outer product in $\mathbb{R}^n$, i.e., $\vec{\boldsymbol{u}} \otimes \vec{\boldsymbol{v}} = \vec{\boldsymbol{u}}\vec{\boldsymbol{v}}^T, \forall \vec{\boldsymbol{u}}, \vec{\boldsymbol{v}} \in \mathbb{R}^n$. In our paper, these two expressions $u \otimes v$ and $uv^T, \forall u, v \in \mathcal{H}$, have the same meaning.*

Proof. $\{\lambda_k(G^*G)\}_{k=1}^r$ and $\{\varphi_k(G^*G)\}_{k=1}^r$ are the respective eigenvalues and eigenvectors of G^*G. Then $G^*G\varphi_k(G^*G) = \lambda_k(G^*G)\varphi_k(G^*G)$. After applying the operator G on both sides, we obtain $GG^*G\varphi_k(G^*G) = \lambda_k(G^*G)G\varphi_k(G^*G)$. So $\{\lambda_k(G^*G)\}_{k=1}^r$ are the positive eigenvalues of GG^*.

To show that $\{G\varphi_k(G^*G)\}_{k=1}^r$ are eigenvectors of GG^*, we need to show that $G\varphi_k(G^*G)$ is nonzero, $\forall k = 1, 2, ...r$. Suppose $G\varphi_k(G^*G) = 0$, then $G^*G\varphi_k(G^*G) = \lambda_k(G^*G)\varphi_k(G^*G) = 0$, which implies that $\varphi_k(G^*G) = 0$, contradicting the fact that $\varphi_k(G^*G)$ is an eigenvector. So $\{G\varphi_k(G^*G)\}_{k=1}^r$ are the eigenvectors of GG^*.

We also need to show $\{\lambda_k(G^*G)\}_{k=1}^r$ are the whole positive eigenvalues of GG^*. It is equivalent to showing that if λ is an positive eigenvalue of GG^*, then λ is also an eigenvalue of G^*G. To achieve this, we can just repeat the above procedure.

Finally, we need to show that $\{\frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}\}_{k=1}^r$ are orthonormal. For any $k, l = 1, 2, ..., r$,

$$\begin{aligned}
\langle \frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}, \frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}} \rangle_{\mathcal{H}_2} &= \frac{1}{\lambda_k(G^*G)} \langle \varphi_k(G^*G), G^*G\varphi_k(G^*G) \rangle_{\mathcal{H}_1} \\
&= \frac{1}{\lambda_k(G^*G)} \langle \varphi_k(G^*G), \lambda_k(G^*G)\varphi_k(G^*G) \rangle_{\mathcal{H}_1} = 1, \\
\langle \frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}, \frac{G\varphi_l(G^*G)}{\sqrt{\lambda_l(G^*G)}} \rangle_{\mathcal{H}_2} &= \langle \frac{\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}, \frac{G^*G\varphi_l(G^*G)}{\sqrt{\lambda_l(G^*G)}} \rangle_{\mathcal{H}_1} = \lambda_l(G^*G) \langle \frac{\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}, \frac{\varphi_l(G^*G)}{\sqrt{\lambda_l(G^*G)}} \rangle_{\mathcal{H}_1} = 0.
\end{aligned}$$

□

Corollary 1. *The projection operator P_{GG^*} on the subspace $\mathrm{Im}(GG^*)$ is*

$$\begin{aligned}
P_{GG^*} &= \sum_{k=1}^r \big[\frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}\big] \otimes \big[\frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}\big] \\
&= G\big[\sum_{k=1}^r \lambda_k(G^*G)^{-1}\varphi_k(G^*G) \otimes \varphi_k(G^*G)\big]G^* \\
&= G(G^*G)^\dagger G^*.
\end{aligned} \tag{3}$$

Proof. From Lemma 2, we have that $\{\frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}\}_{k=1}^r$ are the orthonormal basis of $\mathrm{Im}(GG^*)$. We can obtain P_{GG^*} immediately. □

Corollary 2. *The square root of the operator GG^* is*

$$\begin{aligned}
(GG^*)^{\frac{1}{2}} &= \sum_{k=1}^r \big[\lambda_k(G^*G)\big]^{\frac{1}{2}} \big[\frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}\big] \otimes \big[\frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}\big] \\
&= G\big[\sum_{k=1}^r \lambda_k(G^*G)^{-\frac{1}{2}}\varphi_k(G^*G) \otimes \varphi_k(G^*G)\big]G^* \\
&= G(G^*G)^{\dagger\frac{1}{2}} G^*.
\end{aligned} \tag{4}$$

Proof. Immediately by Lemma 2. □

Corollary 3. *The Moore-Penrose inverse of the square root of the operator GG^* is*

$$\begin{aligned}
(GG^*)^{\dagger\frac{1}{2}} &= \sum_{k=1}^r \big[\lambda_k(G^*G)\big]^{-\frac{1}{2}} \big[\frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}\big] \otimes \big[\frac{G\varphi_k(G^*G)}{\sqrt{\lambda_k(G^*G)}}\big] \\
&= G\big[\sum_{k=1}^r \lambda_k(G^*G)^{-\frac{3}{2}}\varphi_k(G^*G) \otimes \varphi_k(G^*G)\big]G^* \\
&= G(G^*G)^{\dagger\frac{3}{2}} G^*.
\end{aligned} \tag{5}$$

Table 2: The list of notations in the paper and the supplementary material

RKHS	$\mathcal{H}_{\mathcal{K}}$	centered kernel matrix $\boldsymbol{C}_{XX}$	$\boldsymbol{C}_{XX} = \boldsymbol{J}_{N_s}^T \boldsymbol{K}_{XX} \boldsymbol{J}_{N_s}$
kernel function	k	centered kernel matrix $\boldsymbol{C}_{YY}$	$\boldsymbol{C}_{YY} = \boldsymbol{J}_{N_t}^T \boldsymbol{K}_{YY} \boldsymbol{J}_{N_t}$
implicit feature map	ϕ	centered kernel matrix $\boldsymbol{C}_{YX}$	$\boldsymbol{C}_{YX} = \boldsymbol{J}_{N_t}^T \boldsymbol{K}_{YX} \boldsymbol{J}_{N_s}$
identity operator in $\mathcal{H}_{\mathcal{K}}$	$\mathcal{I}_{\mathcal{H}_{\mathcal{K}}}$	top d eigenvalues and eigenvectors of $\boldsymbol{C}_{XX}$	$(\boldsymbol{\Lambda}_X, \boldsymbol{V}_X)$
source samples number	N_s	top d eigenvalues and eigenvectors of $\boldsymbol{C}_{YY}$	$(\boldsymbol{\Lambda}_Y, \boldsymbol{V}_Y)$
target samples number	N_t	matrix $\boldsymbol{W}_X$	$\boldsymbol{W}_X = \boldsymbol{J}_{N_s} \boldsymbol{V}_X (\boldsymbol{I}_d - \rho \boldsymbol{\Lambda}_X^{-1})^{\frac{1}{2}}$
source samples in $\mathcal{H}_{\mathcal{K}}$	Φ_X	matrix $\boldsymbol{W}_Y$	$\boldsymbol{W}_Y = \boldsymbol{J}_{N_t} \boldsymbol{V}_Y (\boldsymbol{I}_d - \rho \boldsymbol{\Lambda}_Y^{-1})^{\frac{1}{2}}$
target samples in $\mathcal{H}_{\mathcal{K}}$	Φ_Y	matrix $\boldsymbol{C}_{XY}^w$	$\boldsymbol{C}_{XY}^w = \boldsymbol{W}_X^T \boldsymbol{K}_{XY} \boldsymbol{W}_Y$
$N_s \times N_s$ centering matrix	$\boldsymbol{J}_{N_s} = \frac{1}{\sqrt{N_s}}(\boldsymbol{I} - \frac{1}{N_s}\vec{\boldsymbol{1}}\vec{\boldsymbol{1}}^T)$	matrix $\boldsymbol{C}_{YX}^w$	$\boldsymbol{C}_{YX}^w = (\boldsymbol{C}_{XY}^w)^T$
$N_t \times N_t$ centering matrix	$\boldsymbol{J}_{N_t} = \frac{1}{\sqrt{N_t}}(\boldsymbol{I} - \frac{1}{N_t}\vec{\boldsymbol{1}}\vec{\boldsymbol{1}}^T)$	matrix $\boldsymbol{B}$	$\boldsymbol{B} = \boldsymbol{C}_{YX}(\boldsymbol{C}_{XX} + \rho \boldsymbol{I}_{N_s})^{-\frac{1}{2}}$
kernel matrix $\boldsymbol{K}_{XX}$	$\boldsymbol{K}_{XX} = \Phi_X^T \Phi_X$	matrix $\boldsymbol{D}$	$\boldsymbol{D} = \left[\boldsymbol{C}_{YX}^w \boldsymbol{C}_{XY}^w + \rho(\boldsymbol{\Lambda}_Y - \rho \boldsymbol{I}_d)\right]^{\dagger \frac{1}{2}}$ $\boldsymbol{W}_Y^T \boldsymbol{K}_{YX} \boldsymbol{J}_{N_s}$
kernel matrix $\boldsymbol{K}_{YY}$	$\boldsymbol{K}_{YY} = \Phi_Y^T \Phi_Y$		
kernel matrix $\boldsymbol{K}_{YX}$	$\boldsymbol{K}_{YX} = \Phi_Y^T \Phi_X$		

Proof. Immediately by Lemma 2. □

Lemma 3. *Let $\mathcal{H}$ be a separable Hilbert space. Let $\psi_1, \psi_2, ..., \psi_n$ be n elements in $\mathcal{H}$. We define the operator $\Psi : \mathbb{R}^n \to \mathcal{H}$ as $\Psi(\boldsymbol{x}) = \sum_{i=1}^n x_i \psi_i, \forall \boldsymbol{x} \in \mathbb{R}^n$. Then $\Psi^T : \mathcal{H} \to \mathbb{R}^n$ defined as $\Psi^T(u) = [\langle \psi_1, u \rangle_{\mathcal{H}}, \langle \psi_2, u \rangle_{\mathcal{H}}, ..., \langle \psi_n, u \rangle_{\mathcal{H}}]^T, \forall u \in \mathcal{H}$ is the adjoint operator of Ψ, i.e., $\Psi^* = \Psi^T$.*

Proof.

$$\forall \boldsymbol{x} \in \mathbb{R}^n, \forall u \in \mathcal{H}, \quad \langle \Psi(\boldsymbol{x}), u \rangle_{\mathcal{H}} = \sum_{i=1}^n x_i \langle \psi_i, u \rangle_{\mathcal{H}} = \langle \boldsymbol{x}, \Psi^T(u) \rangle. \tag{6}$$

□

Note that if $\mathcal{H}_1$ and $\mathcal{H}_2$ are Euclidean spaces, say $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, then the operator $G : \mathcal{H}_1 \to \mathcal{H}_2$ is just an $n_2 \times n_1$ matrix. All the above conclusions still hold. In the next section, we provide the proofs of all mathematical results in the paper. For convenience, we list the relevant notations in Table 2.

3.1. Proving Theorem 1

In $\mathbb{R}^n$, the whitening-coloring map is $\hat{T}_{\mathrm{WC}} = \boldsymbol{\Sigma}_t^{\frac{1}{2}} (\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger}$.

Theorem 1. *If $\mathrm{Im}(\boldsymbol{\Sigma}_t) \subseteq \mathrm{Im}(\boldsymbol{\Sigma}_s)$, then $\hat{\boldsymbol{T}}_{\mathrm{WC}} \boldsymbol{\Sigma}_s \hat{\boldsymbol{T}}_{\mathrm{WC}}^T = \boldsymbol{\Sigma}_t$.*

Proof. Substituting $\hat{\boldsymbol{T}}_{\mathrm{WC}} = \boldsymbol{\Sigma}_t^{\frac{1}{2}} (\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger}$ into the left part, we have

$$\hat{\boldsymbol{T}}_{\mathrm{WC}} \boldsymbol{\Sigma}_s \hat{\boldsymbol{T}}_{\mathrm{WC}}^T = \left[\boldsymbol{\Sigma}_t^{\frac{1}{2}} (\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger}\right] \boldsymbol{\Sigma}_s \left[\boldsymbol{\Sigma}_t^{\frac{1}{2}} (\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger}\right]^T = \boldsymbol{\Sigma}_t^{\frac{1}{2}} (\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger} \boldsymbol{\Sigma}_s (\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger} \boldsymbol{\Sigma}_t^{\frac{1}{2}} = \boldsymbol{\Sigma}_t^{\frac{1}{2}} \boldsymbol{P}_s \boldsymbol{\Sigma}_t^{\frac{1}{2}} = \boldsymbol{\Sigma}_t, \tag{7}$$

where $\boldsymbol{P}_s$ is the projection matrix onto the image space $\mathrm{Im}(\boldsymbol{\Sigma}_s)$, and the last equality holds because $\mathrm{Im}(\boldsymbol{\Sigma}_t) \subseteq \mathrm{Im}(\boldsymbol{\Sigma}_s)$. □

3.2. Proving the equivalence between two expressions of the optimal transport map

Given two positive definite covariance matrix $\boldsymbol{\Sigma}_t$ and $\boldsymbol{\Sigma}_s$, we have

$$\boldsymbol{\Sigma}_t^{\frac{1}{2}} (\boldsymbol{\Sigma}_t^{\frac{1}{2}} \boldsymbol{\Sigma}_s \boldsymbol{\Sigma}_t^{\frac{1}{2}})^{-\frac{1}{2}} \boldsymbol{\Sigma}_t^{\frac{1}{2}} = \boldsymbol{\Sigma}_s^{-\frac{1}{2}} (\boldsymbol{\Sigma}_s^{\frac{1}{2}} \boldsymbol{\Sigma}_t \boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\frac{1}{2}} \boldsymbol{\Sigma}_s^{-\frac{1}{2}}. \tag{8}$$

Proof. We write $\boldsymbol{\Sigma}_t = \boldsymbol{\Sigma}_t^{\frac{1}{2}}(\boldsymbol{\Sigma}_t^{\frac{1}{2}})^T$, and then substitute it into the right part of (8). We have

$$\begin{aligned}
\boldsymbol{\Sigma}_s^{-\frac{1}{2}}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{-\frac{1}{2}} &= \boldsymbol{\Sigma}_s^{-\frac{1}{2}}\big[(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}})(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}})^T\big]^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{-\frac{1}{2}} && \text{(9a)}\\
&= \boldsymbol{\Sigma}_s^{-\frac{1}{2}}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}})\big[(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}})^T(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}})\big]^{-\frac{1}{2}}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}})^T\boldsymbol{\Sigma}_s^{-\frac{1}{2}} && \text{(9b)}\\
&= \boldsymbol{\Sigma}_s^{-\frac{1}{2}}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_t^{\frac{1}{2}})^{-\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{-\frac{1}{2}} && \text{(9c)}\\
&= \boldsymbol{\Sigma}_t^{\frac{1}{2}}(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_t^{\frac{1}{2}})^{-\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}, && \text{(9d)}
\end{aligned}$$

where (9b) holds because of Corollary 2. □

3.3. Proving Theorem 2

In $\mathbb{R}^n$, the optimal transport map is $\hat{\boldsymbol{T}}_{\mathrm{OT}} = \boldsymbol{\Sigma}_t^{\frac{1}{2}}(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_t^{\frac{1}{2}})^{\dagger\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}$.

Theorem 2. *If* $\mathrm{Ker}(\boldsymbol{\Sigma}_s)\cap\mathrm{Im}(\boldsymbol{\Sigma}_t) = \{\vec{\mathbf{0}}\}$, *then we have* $\hat{\boldsymbol{T}}_{\mathrm{OT}}\boldsymbol{\Sigma}_s\hat{\boldsymbol{T}}_{\mathrm{OT}}^T = \boldsymbol{\Sigma}_t$.

Proof. Let $\boldsymbol{\Sigma}_t = \boldsymbol{V}_{n\times r}\boldsymbol{\Lambda}_{r\times r}\boldsymbol{V}_{n\times r}^T$, where $\boldsymbol{\Lambda}$ is a diagonal matrix whose diagonal terms are the r positive eigenvalues and $\boldsymbol{V}$ consists of the corresponding eigenvectors.

(I) Claim: $\mathrm{rank}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{V}) = r$.

Let $\vec{\boldsymbol{v}}_1, \vec{\boldsymbol{v}}_2, ..., \vec{\boldsymbol{v}}_r$ be the columns of $\boldsymbol{V}$. It is sufficient to show that $\boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_1, \boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_2, ..., \boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_r$ are linearly independent.

Suppose there exist $\lambda_1, \lambda_2, ..., \lambda_r$, such that $\lambda_1\boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_1 + \lambda_2\boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_2 + ... + \lambda_r\boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_r = \vec{\mathbf{0}}$. Then we have that $\lambda_1\vec{\boldsymbol{v}}_1 + \lambda_2\vec{\boldsymbol{v}}_2 + ...+\lambda_r\vec{\boldsymbol{v}}_r \in \mathrm{Ker}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}) = \mathrm{Ker}(\boldsymbol{\Sigma}_s)$. Since $\mathrm{Im}(\boldsymbol{\Sigma}_t) = \mathrm{Span}\{\vec{\boldsymbol{v}}_1, \vec{\boldsymbol{v}}_2, .., \vec{\boldsymbol{v}}_r\}$, immediately, $\lambda_1\vec{\boldsymbol{v}}_1+\lambda_2\vec{\boldsymbol{v}}_2+...+\lambda_r\vec{\boldsymbol{v}}_r \in \mathrm{Im}(\boldsymbol{\Sigma}_t)$. By the condition that $\mathrm{Ker}(\boldsymbol{\Sigma}_s)\cap\mathrm{Im}(\boldsymbol{\Sigma}_t) = \{\vec{\mathbf{0}}\}$, we have $\lambda_1\vec{\boldsymbol{v}}_1 + \lambda_2\vec{\boldsymbol{v}}_2 + ... + \lambda_r\vec{\boldsymbol{v}}_r = \vec{\mathbf{0}}$. And $\vec{\boldsymbol{v}}_1, \vec{\boldsymbol{v}}_2, ..., \vec{\boldsymbol{v}}_r$ are linearly independent $\Longrightarrow\ \lambda_1 = \lambda_2 = ... = \lambda_r = 0$. So $\boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_1, \boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_2, ..., \boldsymbol{\Sigma}_s^{\frac{1}{2}}\vec{\boldsymbol{v}}_r$ are linearly independent.

(II) Now we start to prove that $\hat{\boldsymbol{T}}_{\mathrm{OT}}\boldsymbol{\Sigma}_s\hat{\boldsymbol{T}}_{\mathrm{OT}}^T = \boldsymbol{\Sigma}_t$. Substituting $\hat{\boldsymbol{T}}_{\mathrm{OT}} = \boldsymbol{\Sigma}_t^{\frac{1}{2}}(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_t^{\frac{1}{2}})^{\dagger\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}$ into the left part, we obtain

$$\hat{\boldsymbol{T}}_{\mathrm{OT}}\boldsymbol{\Sigma}_s\hat{\boldsymbol{T}}_{\mathrm{OT}}^T = \boldsymbol{\Sigma}_t^{\frac{1}{2}}(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_t^{\frac{1}{2}})^{\dagger\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_t^{\frac{1}{2}}(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_t^{\frac{1}{2}})^{\dagger\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}} = \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{P}_{ts}\boldsymbol{\Sigma}_t^{\frac{1}{2}}, \tag{10}$$

where $\boldsymbol{P}_{ts}$ is the projection matrix onto $\mathrm{Im}(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s\boldsymbol{\Sigma}_t^{\frac{1}{2}}) = \mathrm{Im}\big[(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{\frac{1}{2}})(\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{\frac{1}{2}})^T\big]$. We set $G = \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{\frac{1}{2}}$, then we obtain $\boldsymbol{P}_{ts} = \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}$ by Corollary 1. So,

$$\hat{\boldsymbol{T}}_{\mathrm{OT}}\boldsymbol{\Sigma}_s\hat{\boldsymbol{T}}_{\mathrm{OT}}^T = \boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}}\boldsymbol{\Sigma}_t^{\frac{1}{2}} = \boldsymbol{\Sigma}_t\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{\Sigma}_t. \tag{11}$$

Substituting $\boldsymbol{\Sigma}_t = \boldsymbol{V}_{n\times r}\boldsymbol{\Lambda}_{r\times r}\boldsymbol{V}_{n\times r}^T$ into the above, we get

$$\begin{aligned}
\hat{\boldsymbol{T}}_{\mathrm{OT}}\boldsymbol{\Sigma}_s\hat{\boldsymbol{T}}_{\mathrm{OT}}^T &= \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}}(\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}})^{\dagger}\boldsymbol{\Sigma}_s^{\frac{1}{2}}\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T\\
&= \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}})\big[(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}})^T(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}})\big]^{\dagger}(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}})^T\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\\
&= \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{P}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T,
\end{aligned} \tag{12}$$

where $\boldsymbol{P}$ is the projection matrix onto $\mathrm{Im}(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s\boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}) = \mathrm{Im}(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}})$, and the last equality in (12) holds because of Corollary 1. Note that $\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}}$ is an $r\times n$ matrix, so $\mathrm{Im}(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}}) \subseteq \mathbb{R}^r$. In addition, by the above claim, $\mathrm{rank}(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}}) = \mathrm{rank}(\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}}) = r$, so we have that $\mathrm{Im}(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T\boldsymbol{\Sigma}_s^{\frac{1}{2}}) = \mathbb{R}^r$. Therefore, the projection matrix $\boldsymbol{P}$ is just the identity matrix $\boldsymbol{I}_{r\times r}$. Finally, Eq (12) can be written as $\hat{\boldsymbol{T}}_{\mathrm{OT}}\boldsymbol{\Sigma}_s\hat{\boldsymbol{T}}_{\mathrm{OT}}^T = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{I}_{r\times r}\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T = \boldsymbol{\Sigma}_t$, and we get the desired conclusion. □

3.4. Proving Proposition 1

The maximum likelihood estimations of source and target covariance descriptors are given by

$$\mathrm{MC}_s = (\Phi_X \boldsymbol{J}_{N_s})(\Phi_X \boldsymbol{J}_{N_s})^T + \rho I_{\mathcal{H}_\mathcal{K}} \tag{13a}$$

$$\mathrm{MC}_t = (\Phi_Y \boldsymbol{J}_{N_t})(\Phi_Y \boldsymbol{J}_{N_t})^T. \tag{13b}$$

Proposition 1. *With the maximum likelihood estimators (13), the kernel whitening-coloring map is given by*

$$\mathrm{k}\hat{T}_{\mathrm{WC}} = (\mathrm{MC}_t)^{\frac{1}{2}}(\mathrm{MC}_s)^{\dagger\frac{1}{2}} = \Phi_Y \boldsymbol{J}_{N_t} \boldsymbol{C}_{YY}^{\dagger\frac{1}{2}} \big[\boldsymbol{C}_{YX} \boldsymbol{A} \boldsymbol{J}_{N_s} \Phi_X^T + \frac{1}{\sqrt{\rho}} \boldsymbol{J}_{N_t} \Phi_Y^T\big], \tag{14}$$

where $\boldsymbol{C}_{YY} = \boldsymbol{J}_{N_t}^T \boldsymbol{K}_{YY} \boldsymbol{J}_{N_t}$ *and* $\boldsymbol{C}_{YX} = \boldsymbol{J}_{N_t}^T \boldsymbol{K}_{YX} \boldsymbol{J}_{N_s}$ *are centered kernel matrices, and* $\boldsymbol{A} = \sum_{k=1}^r \frac{1}{\lambda_k}(\frac{1}{\sqrt{\lambda_k+\rho}} - \frac{1}{\sqrt{\rho}})\vec{\boldsymbol{v}}_k \vec{\boldsymbol{v}}_k^T$, *and* $\{\lambda_k, \vec{\boldsymbol{v}}_k\}_{k=1}^r$ *are positive eigenpairs of* $\boldsymbol{C}_{XX}$.

Proof. We substitute maximum likelihood estimators (13) into the expressions of $\mathrm{k}\hat{T}_{\mathrm{WC}}$, and then we have,

$$\mathrm{k}\hat{T}_{\mathrm{WC}} = \big[(\Phi_Y \boldsymbol{J}_{N_t})(\Phi_Y \boldsymbol{J}_{N_t})^T\big]^{\frac{1}{2}} \big[(\Phi_X \boldsymbol{J}_{N_s})(\Phi_X \boldsymbol{J}_{N_s})^T + \rho I_{\mathcal{H}_\mathcal{K}}\big]^{-\frac{1}{2}}. \tag{15}$$

For simplicity, we set $G = \Phi_X \boldsymbol{J}_{N_s}$, then by Lemma 3, $G^* = (\Phi_X \boldsymbol{J}_{N_s})^T$. Let $\{\lambda_k\}_{k=1}^r$ and $\{\vec{\boldsymbol{v}}_k\}_{k=1}^r$ be the positive eigenvalues and the corresponding orthonormal eigenvectors of $G^*G = (\Phi_X \boldsymbol{J}_{N_s})^T(\Phi_X \boldsymbol{J}_{N_s}) = \boldsymbol{C}_{XX}$. Then by Lemma 2, $\{\lambda_k, \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}}\}_{k=1}^r$ are the positive eigenpairs of GG^*. Let $\{\psi_k\}_{k=r+1}^{\dim(\mathcal{H}_\mathcal{K})}$ be a set of orthonormal vectors, such that

$$\{\frac{G\vec{\boldsymbol{v}}_1}{\sqrt{\lambda_1}}, \frac{G\vec{\boldsymbol{v}}_2}{\sqrt{\lambda_2}}, ..., \frac{G\vec{\boldsymbol{v}}_r}{\sqrt{\lambda_r}}, \psi_{r+1}, \psi_{r+2}, ...\} \tag{16}$$

is a complete orthonormal basis of $\mathcal{H}_\mathcal{K}$. Then the identity operator $I_{\mathcal{H}_\mathcal{K}}$ can be written as $I_{\mathcal{H}_\mathcal{K}} = \sum_{k=1}^r \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} \otimes \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} + \sum_{k=r+1}^{\dim(\mathcal{H}_\mathcal{K})} \psi_k \otimes \psi_k$. Therefore, we have

$$\big[(\Phi_X \boldsymbol{J}_{N_s})(\Phi_X \boldsymbol{J}_{N_s})^T + \rho I_{\mathcal{H}_\mathcal{K}}\big]^{-\frac{1}{2}} \tag{17a}$$

$$= (GG^* + \rho I_{\mathcal{H}_\mathcal{K}})^{-\frac{1}{2}} \tag{17b}$$

$$= \big[\sum_{k=1}^r \lambda_k \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} \otimes \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} + \rho I_{\mathcal{H}_\mathcal{K}}\big]^{-\frac{1}{2}} \tag{17c}$$

$$= \big[\sum_{k=1}^r (\lambda_k + \rho) \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} \otimes \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} + \sum_{k=r+1}^{\dim(\mathcal{H}_\mathcal{K})} \rho \psi_k \otimes \psi_k\big]^{-\frac{1}{2}} \tag{17d}$$

$$= \sum_{k=1}^r \frac{1}{\sqrt{\lambda_k + \rho}} \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} \otimes \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} + \sum_{k=r+1}^{\dim(\mathcal{H}_\mathcal{K})} \frac{1}{\sqrt{\rho}} \psi_k \otimes \psi_k \tag{17e}$$

$$= \sum_{k=1}^r \frac{1}{\sqrt{\lambda_k + \rho}} \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} \otimes \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} + \frac{1}{\sqrt{\rho}} (I_{\mathcal{H}_\mathcal{K}} - \sum_{k=1}^r \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}} \otimes \frac{G\vec{\boldsymbol{v}}_k}{\sqrt{\lambda_k}}) \tag{17f}$$

$$= G(\sum_{k=1}^r \frac{1}{\lambda_k \sqrt{\lambda_k + \rho}} \vec{\boldsymbol{v}}_k \vec{\boldsymbol{v}}_k^T) G^* + \frac{1}{\sqrt{\rho}} \big[I_{\mathcal{H}_\mathcal{K}} - G(\sum_{k=1}^r \frac{1}{\lambda_k} \vec{\boldsymbol{v}}_k \vec{\boldsymbol{v}}_k^T) G^*\big] \tag{17g}$$

$$= GAG^* + \frac{1}{\sqrt{\rho}} I_{\mathcal{H}_\mathcal{K}} \tag{17h}$$

$$= \Phi_X \boldsymbol{J}_{N_s} \boldsymbol{A} \boldsymbol{J}_{N_s}^T \Phi_X^T + \frac{1}{\sqrt{\rho}} I_{\mathcal{H}_\mathcal{K}}. \tag{17i}$$

Substituting (17i) into (15), we get

$$\mathrm{k}\hat{T}_{\mathrm{WC}} = \left[(\Phi_Y \boldsymbol{J}_{N_t})(\Phi_Y \boldsymbol{J}_{N_t})^T\right]^{\frac{1}{2}} \left[(\Phi_X \boldsymbol{J}_{N_s})(\Phi_X \boldsymbol{J}_{N_s})^T + \rho I_{\mathcal{H}_\mathcal{K}}\right]^{-\frac{1}{2}} \tag{18a}$$

$$= \left[(\Phi_Y \boldsymbol{J}_{N_t})(\Phi_Y \boldsymbol{J}_{N_t})^T\right]^{\frac{1}{2}} (\Phi_X \boldsymbol{J}_{N_s} \boldsymbol{A} \boldsymbol{J}_{N_s}^T \Phi_X^T + \frac{1}{\sqrt{\rho}} I_{\mathcal{H}_\mathcal{K}}) \tag{18b}$$

$$= \Phi_Y \boldsymbol{J}_{N_t} \left[\boldsymbol{J}_{N_t}^T \Phi_Y^T \Phi_Y \boldsymbol{J}_{N_t}\right]^{\dagger\frac{1}{2}} \boldsymbol{J}_{N_t}^T \Phi_Y^T (\Phi_X \boldsymbol{J}_{N_s} \boldsymbol{A} \boldsymbol{J}_{N_s}^T \Phi_X^T + \frac{1}{\sqrt{\rho}} I_{\mathcal{H}_\mathcal{K}}) \tag{18c}$$

$$= \Phi_Y \boldsymbol{J}_{N_t} \boldsymbol{C}_{YY}^{\dagger\frac{1}{2}} \left[\boldsymbol{C}_{YX} \boldsymbol{A} \boldsymbol{J}_{N_s} \Phi_X^T + \frac{1}{\sqrt{\rho}} \boldsymbol{J}_{N_t} \Phi_Y^T\right], \tag{18d}$$

where (18c) holds because of Corollary 2. □

3.5. Proving Proposition 2

The computationally efficient estimations of source and target covariance descriptors are given by

$$\mathrm{E}C_s = (\Phi_X \boldsymbol{W}_X)(\Phi_X \boldsymbol{W}_X)^T + \rho I_{\mathcal{H}_\mathcal{K}} \tag{19a}$$

$$\mathrm{E}C_t = (\Phi_Y \boldsymbol{W}_Y)(\Phi_Y \boldsymbol{W}_Y)^T. \tag{19b}$$

Proposition 2. *With the computationally efficient estimators (19), the kernel optimal transport map is given by*

$$\mathrm{k}\hat{T}_{\mathrm{OT}} = (\mathrm{E}C_t)^{\frac{1}{2}} \left[(\mathrm{E}C_t)^{\frac{1}{2}} (\mathrm{E}C_s)(\mathrm{E}C_t)^{\frac{1}{2}}\right]^{\dagger\frac{1}{2}} (\mathrm{E}C_t)^{\frac{1}{2}} = \Phi_Y \boldsymbol{W}_Y \left[\boldsymbol{C}_{YX}^w \boldsymbol{C}_{XY}^w + \rho(\boldsymbol{\Lambda}_Y - \rho \boldsymbol{I}_d)\right]^{\dagger\frac{1}{2}} \boldsymbol{W}_Y^T \Phi_Y^T, \tag{20}$$

where $\boldsymbol{C}_{YX}^w = \boldsymbol{W}_Y^T \boldsymbol{K}_{YX} \boldsymbol{W}_X$ and $\boldsymbol{C}_{XY}^w = (\boldsymbol{C}_{YX}^w)^T$, and $\boldsymbol{\Lambda}_Y$ is the diagonal matrix storing the top d eigenvalues of $\boldsymbol{C}_{YY}$.

Proof. We substitute the computationally efficient estimators (19) into the expression of $\mathrm{k}\hat{T}_{\mathrm{OT}}$, and then we have,

$$\mathrm{k}\hat{T}_{\mathrm{OT}} = (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \left[(\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} (\Phi_X \boldsymbol{W}_X \boldsymbol{W}_X^T \Phi_X^T + \rho I_{\mathcal{H}_\mathcal{K}}) (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}}\right]^{\dagger\frac{1}{2}} (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \tag{21}$$

We observe that $\mathrm{E}C_s = \Phi_X \boldsymbol{W}_X \boldsymbol{W}_X^T \Phi_X^T + \rho I_{\mathcal{H}_\mathcal{K}}$ is strictly positive definite, which implies that there exists an operator $\mathcal{A}$ (*e.g.*, $\mathcal{A} = (\mathrm{E}C_s)^{\frac{1}{2}}$), such that $\mathcal{A}\mathcal{A}^T = \mathrm{E}C_s$. Then $\mathrm{k}\hat{T}_{\mathrm{OT}}$ can be written as

$$\mathrm{k}\hat{T}_{\mathrm{OT}} = (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \left[(\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} (\mathcal{A}\mathcal{A}^T) (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}}\right]^{\dagger\frac{1}{2}} (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \tag{22a}$$

$$= (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \left(\left[(\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \mathcal{A}\right] \left[(\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \mathcal{A}\right]^T\right)^{\dagger\frac{1}{2}} (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \tag{22b}$$

$$= (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \left[(\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \mathcal{A}\right] \left(\left[(\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \mathcal{A}\right]^T \left[(\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \mathcal{A}\right]\right)^{\dagger\frac{3}{2}} \tag{22c}$$

$$\left[(\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \mathcal{A}\right]^T (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T)^{\frac{1}{2}} \tag{22d}$$

$$= (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T) \mathcal{A} \left[\mathcal{A}^T \Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T \mathcal{A}\right]^{\dagger\frac{3}{2}} \mathcal{A}^T (\Phi_Y \boldsymbol{W}_Y \boldsymbol{W}_Y^T \Phi_Y^T) \tag{22e}$$

$$= \Phi_Y \boldsymbol{W}_Y (\boldsymbol{W}_Y^T \Phi_Y^T \mathcal{A}) \left[(\boldsymbol{W}_Y^T \Phi_Y^T \mathcal{A})^T (\boldsymbol{W}_Y^T \Phi_Y^T \mathcal{A})\right]^{\dagger\frac{3}{2}} (\boldsymbol{W}_Y^T \Phi_Y^T \mathcal{A})^T \boldsymbol{W}_Y^T \Phi_Y^T \tag{22f}$$

$$= \Phi_Y \boldsymbol{W}_Y \left[(\boldsymbol{W}_Y^T \Phi_Y^T \mathcal{A})(\boldsymbol{W}_Y^T \Phi_Y^T \mathcal{A})^T\right]^{\dagger\frac{1}{2}} \boldsymbol{W}_Y^T \Phi_Y^T \tag{22g}$$

$$= \Phi_Y \boldsymbol{W}_Y (\boldsymbol{W}_Y^T \Phi_Y^T \mathcal{A}\mathcal{A}^T \Phi_Y \boldsymbol{W}_Y)^{\dagger\frac{1}{2}} \boldsymbol{W}_Y^T \Phi_Y^T \tag{22h}$$

$$= \Phi_Y \boldsymbol{W}_Y \left[\boldsymbol{W}_Y^T \Phi_Y^T (\Phi_X \boldsymbol{W}_X \boldsymbol{W}_X^T \Phi_X^T + \rho I_{\mathcal{H}_\mathcal{K}}) \Phi_Y \boldsymbol{W}_Y\right]^{\dagger\frac{1}{2}} \boldsymbol{W}_Y^T \Phi_Y^T \tag{22i}$$

$$= \Phi_Y \boldsymbol{W}_Y (\boldsymbol{C}_{YX}^w \boldsymbol{C}_{XY}^w + \rho \boldsymbol{W}_Y^T \boldsymbol{K}_{YY} \boldsymbol{W}_Y)^{\dagger\frac{1}{2}} \boldsymbol{W}_Y^T \Phi_Y^T, \tag{22j}$$

where (22c) and (22g) hold because of Corollary 3. Next we prove that $\boldsymbol{W}_Y^T \boldsymbol{K}_{YY} \boldsymbol{W}_Y = \boldsymbol{\Lambda}_Y - \rho \boldsymbol{I}_d$:

$$\begin{aligned}
\boldsymbol{W}_Y^T \boldsymbol{K}_{YY} \boldsymbol{W}_Y &= \left[\boldsymbol{J}_{N_t} \boldsymbol{V}_Y (\boldsymbol{I}_d - \rho \boldsymbol{\Lambda}_{\boldsymbol{Y}}^{-1})^{\frac{1}{2}}\right]^T \boldsymbol{K}_{YY} \left[\boldsymbol{J}_{N_t} \boldsymbol{V}_Y (\boldsymbol{I}_d - \rho \boldsymbol{\Lambda}_{\boldsymbol{Y}}^{-1})^{\frac{1}{2}}\right] \\
&= (\boldsymbol{I}_d - \rho \boldsymbol{\Lambda}_{\boldsymbol{Y}}^{-1})^{\frac{1}{2}} \boldsymbol{V}_Y^T \boldsymbol{C}_{YY} \boldsymbol{V}_Y^T (\boldsymbol{I}_d - \rho \boldsymbol{\Lambda}_{\boldsymbol{Y}}^{-1})^{\frac{1}{2}} \\
&= (\boldsymbol{I}_d - \rho \boldsymbol{\Lambda}_{\boldsymbol{Y}}^{-1})^{\frac{1}{2}} \boldsymbol{\Lambda}_Y (\boldsymbol{I}_d - \rho \boldsymbol{\Lambda}_{\boldsymbol{Y}}^{-1})^{\frac{1}{2}} \\
&= \boldsymbol{\Lambda}_Y - \rho \boldsymbol{I}_d.
\end{aligned} \tag{23}$$

After substituting (23) into (22j), we obtain the desired result. □

3.6. Derivations of the domain-invariant kernel expressions

The domain-invariant kernel $\triangle \tilde{\boldsymbol{K}}$ is given by

$$\triangle \tilde{\boldsymbol{K}} = \begin{bmatrix} \triangle \tilde{\boldsymbol{K}}_{ss} & \triangle \tilde{\boldsymbol{K}}_{ts}^T \\ \triangle \tilde{\boldsymbol{K}}_{ts} & \triangle \tilde{\boldsymbol{K}}_{tt} \end{bmatrix} = \begin{bmatrix} \Psi_{s\to t}^T \Psi_{s\to t} & \Psi_{s\to t}^T \Psi_t \\ \Psi_t^T \Psi_{s\to t} & \Psi_t^T \Psi_t \end{bmatrix}, \tag{24}$$

where the symbol $\triangle$ represents the way of "moving" the source samples, *i.e.*, $\triangle$ = WC or OT, and $\Psi_{s\to t}$ denotes the transported source samples, *i.e.*, $\Psi_{s\to t} = \mathrm{k}T_{\triangle}(\Psi_s)$, and Ψ_s and Ψ_t denote the centered source samples and target samples, respectively.

Using the kernel whitening-coloring map (14), we get

$$\begin{aligned}
\Psi_{s\to t} &= \mathrm{k}\hat{T}_{\mathrm{WC}}(\Psi_s) = \sqrt{N_s} \Phi_Y \boldsymbol{J}_{N_t} \boldsymbol{C}_{YY}^{\dagger \frac{1}{2}} \boldsymbol{B} \\
\mathrm{WC}\tilde{\boldsymbol{K}}_{ss} &= N_s \boldsymbol{B}^T \boldsymbol{B} \\
\mathrm{WC}\tilde{\boldsymbol{K}}_{ts} &= \sqrt{N_s N_t} \boldsymbol{C}_{YY}^{\frac{1}{2}} \boldsymbol{B} = \sqrt{N_s N_t} \boldsymbol{U}_Y \boldsymbol{\Lambda}_Y^{\frac{1}{2}} \boldsymbol{U}_Y^T \boldsymbol{B},
\end{aligned} \tag{25}$$

where $\boldsymbol{B} = \boldsymbol{C}_{YX} (\boldsymbol{C}_{XX} + \rho \boldsymbol{I}_{N_s})^{-\frac{1}{2}}$, and $(\boldsymbol{U}_Y, \Lambda_Y^{\frac{1}{2}})$ stores the top d eigenpairs of $\boldsymbol{C}_{YY}$. Note that, in practice, in order to exploit the principal components and reduce the computational complexity, we artificially select d to be a small integer, *i.e.*, $d \ll N_t$.

Proof.
(I) The transported source samples are

$$\Psi_{s\to t} = \mathrm{k}T_{\mathrm{WC}}(\Psi_s) \tag{26a}$$

$$= \Phi_Y \boldsymbol{J}_{N_t} \boldsymbol{C}_{YY}^{\dagger \frac{1}{2}} \left[\boldsymbol{C}_{YX} \boldsymbol{A} \boldsymbol{J}_{N_s} \Phi_X^T + \frac{1}{\sqrt{\rho}} \boldsymbol{J}_{N_t} \Phi_Y^T\right] (\sqrt{N_s} \Phi_X \boldsymbol{J}_{N_s}) \tag{26b}$$

$$= \sqrt{N_s} \Phi_Y \boldsymbol{J}_{N_t} \boldsymbol{C}_{YY}^{\dagger \frac{1}{2}} \boldsymbol{C}_{YX} (\boldsymbol{A} \boldsymbol{C}_{XX} + \frac{1}{\sqrt{\rho}} \boldsymbol{I}_{N_s}). \tag{26c}$$

Now we consider the term $\boldsymbol{A}\boldsymbol{C}_{XX} + \frac{1}{\sqrt{\rho}} \boldsymbol{I}_{N_s}$. Recall that $\{\vec{\boldsymbol{v}}_1, \vec{\boldsymbol{v}}_2, ..., \vec{\boldsymbol{v}}_r\}$ are $\boldsymbol{C}_{XX}$'s eigenvectors, the corresponding eigenvalues of which are positive. Let $\{\vec{\boldsymbol{u}}_{r+1}, \vec{\boldsymbol{u}}_{r+2}, ..., \vec{\boldsymbol{u}}_{N_s}\}$ be a set of orthonormal vectors, such that $\{\vec{\boldsymbol{v}}_1, \vec{\boldsymbol{v}}_2, ..., \vec{\boldsymbol{v}}_r, \vec{\boldsymbol{u}}_{r+1}, \vec{\boldsymbol{u}}_{r+2},$

$\ldots, \vec{\boldsymbol{u}}_{N_s}\}$ is an orthonormal system for $\mathbb{R}^{N_s}$. Then, we have

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{C}_{XX} + \frac{1}{\sqrt{\rho}}\boldsymbol{I}_{N_s} =& \sum_{k=1}^{r} \frac{1}{\lambda_k}\Big(\frac{1}{\sqrt{\lambda_k+\rho}} - \frac{1}{\sqrt{\rho}}\Big)\vec{\boldsymbol{v}}_k\vec{\boldsymbol{v}}_k^T\boldsymbol{C}_{XX} + \frac{1}{\sqrt{\rho}}\boldsymbol{I}_{N_s} && \text{(27a)}\\
=& \sum_{k=1}^{r} \frac{1}{\lambda_k}\Big(\frac{1}{\sqrt{\lambda_k+\rho}} - \frac{1}{\sqrt{\rho}}\Big)\vec{\boldsymbol{v}}_k(\boldsymbol{C}_{XX}\vec{\boldsymbol{v}}_k)^T + \frac{1}{\sqrt{\rho}}\boldsymbol{I}_{N_s} && \text{(27b)}\\
=& \sum_{k=1}^{r}\Big(\frac{1}{\sqrt{\lambda_k+\rho}} - \frac{1}{\sqrt{\rho}}\Big)\vec{\boldsymbol{v}}_k\vec{\boldsymbol{v}}_k^T + \frac{1}{\sqrt{\rho}}\sum_{k=1}^{r}\vec{\boldsymbol{v}}_k\vec{\boldsymbol{v}}_k^T + \frac{1}{\sqrt{\rho}}\sum_{k=r+1}^{N_s}\vec{\boldsymbol{u}}_k\vec{\boldsymbol{u}}_k^T && \text{(27c)}\\
=& \sum_{k=1}^{r}\frac{1}{\sqrt{\lambda_k+\rho}}\vec{\boldsymbol{v}}_k\vec{\boldsymbol{v}}_k^T + \frac{1}{\sqrt{\rho}}\sum_{k=r+1}^{N_s}\vec{\boldsymbol{u}}_k\vec{\boldsymbol{u}}_k^T && \text{(27d)}\\
=& \Big[\sum_{k=1}^{r}(\lambda_k+\rho)\vec{\boldsymbol{v}}_k\vec{\boldsymbol{v}}_k^T + \sum_{k=r+1}^{N_s}\rho\vec{\boldsymbol{u}}_k\vec{\boldsymbol{u}}_k^T\Big]^{-\frac{1}{2}} && \text{(27e)}\\
=& (\boldsymbol{C}_{XX}+\rho\boldsymbol{I}_{N_s})^{-\frac{1}{2}}. && \text{(27f)}
\end{aligned}$$

Write $\boldsymbol{B} = \boldsymbol{C}_{YX}(\boldsymbol{C}_{XX}+\rho\boldsymbol{I}_{N_s})^{-\frac{1}{2}}$, then $\Psi_{s\to t} = \mathrm{k}\hat{T}_{\mathrm{WC}}(\Psi_s) = \sqrt{N_s}\Phi_Y\boldsymbol{J}_{N_t}\boldsymbol{C}_{YY}^{\dagger\frac{1}{2}}\boldsymbol{B}$.

(II)

$$\begin{aligned}
\mathrm{WC}\tilde{\boldsymbol{K}}_{ss} &= \Psi_{s\to t}^T\Psi_{s\to t} = (\sqrt{N_s}\Phi_Y\boldsymbol{J}_{N_t}\boldsymbol{C}_{YY}^{\dagger\frac{1}{2}}\boldsymbol{B})^T(\sqrt{N_s}\Phi_Y\boldsymbol{J}_{N_t}\boldsymbol{C}_{YY}^{\dagger\frac{1}{2}}\boldsymbol{B})\\
&= N_s\boldsymbol{B}^T\boldsymbol{C}_{YY}^{\dagger\frac{1}{2}}\boldsymbol{J}_{N_t}^T\Phi_Y^T\Phi_Y\boldsymbol{J}_{N_t}\boldsymbol{C}_{YY}^{\dagger\frac{1}{2}}\boldsymbol{B} = N_s\boldsymbol{B}^T\boldsymbol{P}_{YY}\boldsymbol{B} = N_s\boldsymbol{B}^T\boldsymbol{B},
\end{aligned} \tag{28}$$

where $\boldsymbol{P}_{YY}$ is the projection matrix onto $\mathrm{Im}(\boldsymbol{C}_{YY}) = \mathrm{Im}(\boldsymbol{J}_{N_t}^T\Phi_Y^T)$, and the last equality holds because $\mathrm{Im}(\boldsymbol{B}) \subseteq \mathrm{Im}(\boldsymbol{C}_{YX}) \subseteq \mathrm{Im}(\boldsymbol{J}_{N_t}^T\Phi_Y^T)$.

(III)

$$\mathrm{WC}\tilde{\boldsymbol{K}}_{ts} = \Psi_t^T\Psi_{s\to t} = (\sqrt{N_t}\Phi_Y\boldsymbol{J}_{N_t})^T(\sqrt{N_s}\Phi_Y\boldsymbol{J}_{N_t}\boldsymbol{C}_{YY}^{\dagger\frac{1}{2}}\boldsymbol{B}) = \sqrt{N_tN_s}\boldsymbol{C}_{YY}^{\dagger\frac{1}{2}}\boldsymbol{B}. \tag{29}$$

□

Using the kernel optimal transport map (20), we get

$$\begin{aligned}
\Psi_{s\to t} &= \mathrm{k}\hat{T}_{\mathrm{OT}}(\Psi_s) = \sqrt{N_s}\Phi_Y\boldsymbol{W}_Y\boldsymbol{D}\\
\mathrm{OT}\tilde{\boldsymbol{K}}_{ss} &= N_s\boldsymbol{D}^T(\boldsymbol{\Lambda}_Y - \rho\boldsymbol{I}_d)\boldsymbol{D}\\
\mathrm{OT}\tilde{\boldsymbol{K}}_{ts} &= \sqrt{N_sN_t}\boldsymbol{J}_{N_t}\boldsymbol{K}_{YY}\boldsymbol{W}_Y\boldsymbol{D},
\end{aligned} \tag{30}$$

where $\boldsymbol{D} = \big[\boldsymbol{C}_{YX}^w\boldsymbol{C}_{XY}^w + \rho(\boldsymbol{\Lambda}_Y - \rho\boldsymbol{I}_d)\big]^{\dagger\frac{1}{2}}\boldsymbol{W}_Y^T\boldsymbol{K}_{YX}\boldsymbol{J}_{N_s}$.

Proof.

(I) The transported samples are

$$\begin{aligned}
\Psi_{s\to t} = \mathrm{k}\hat{T}_{\mathrm{OT}}(\Psi_s) &= \Phi_Y\boldsymbol{W}_Y\big[\boldsymbol{C}_{YX}^w\boldsymbol{C}_{XY}^w + \rho(\boldsymbol{\Lambda}_Y - \rho\boldsymbol{I}_d)\big]^{\dagger\frac{1}{2}}\boldsymbol{W}_Y^T\Phi_Y^T(\sqrt{N_s}\Phi_X\boldsymbol{J}_{N_s})\\
&= \sqrt{N_s}\Phi_Y\boldsymbol{W}_Y\big[\boldsymbol{C}_{YX}^w\boldsymbol{C}_{XY}^w + \rho(\boldsymbol{\Lambda}_Y - \rho\boldsymbol{I}_d)\big]^{\dagger\frac{1}{2}}\boldsymbol{W}_Y^T\boldsymbol{K}_{YX}\boldsymbol{J}_{N_s} = \sqrt{N_s}\Phi_Y\boldsymbol{W}_Y\boldsymbol{D}.
\end{aligned} \tag{31}$$

(II)

$$\mathrm{OT}\tilde{\boldsymbol{K}}_{ss} = \Psi_{s\to t}^T\Psi_{s\to t} = (\sqrt{N_s}\Phi_Y\boldsymbol{W}_Y\boldsymbol{D})^T(\sqrt{N_s}\Phi_Y\boldsymbol{W}_Y\boldsymbol{D}) = N_s\boldsymbol{D}^T\boldsymbol{W}_Y^T\boldsymbol{K}_{YY}\boldsymbol{W}_Y\boldsymbol{D} = N_s\boldsymbol{D}^T(\boldsymbol{\Lambda}_Y - \rho\boldsymbol{I}_d)\boldsymbol{D}, \tag{32}$$

where the last equality holds because of (23).

(III)

$$\mathrm{OT}\tilde{\boldsymbol{K}}_{ts} = \Psi_t^T\Psi_{s\to t} = (\sqrt{N_t}\Phi_Y\boldsymbol{J}_{N_t})^T(\sqrt{N_s}\Phi_Y\boldsymbol{W}_Y\boldsymbol{D}) = \sqrt{N_sN_t}\boldsymbol{J}_{N_t}\boldsymbol{K}_{YY}\boldsymbol{W}_Y\boldsymbol{D}. \tag{33}$$

□

References

[1] M. Harandi, M. Salzmann, and F. Porikli. Bregman divergences for infinite dimensional covariance matrices. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1003–1010, 2014. 1

[2] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks*, 22(2):199–210, 2011. 2

[3] M. H. Quang, M. San Biagio, and V. Murino. Log-hilbert-schmidt metric between positive definite operators on hilbert spaces. In *Advances in Neural Information Processing Systems*, pages 388–396, 2014. 1

[4] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998. 1, 3