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

\[
MC = \frac{1}{N} \sum_{i=1}^{N} [\phi(x_i) - \hat{\mu}] [\phi(x_i) - \hat{\mu}]^T = \Phi X J_N J_N^T \Phi_X^T, \tag{1}
\]

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, \( i.e., k(\vec{x}, \vec{y}) = \vec{x}^T \vec{y} \), then the corresponding feature map is just the identity function, \( i.e., \phi(\vec{x}) = \vec{x} \). 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.hockey. 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 \( 963 + 979 + 987 + 993 = 3922 \)
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.

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

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 $K$. 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 $WC \tilde{K}$ and $OT \tilde{K}$ (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 $K$ and domain-invariant kernel matrices $WC \tilde{K}$ and $OT \tilde{K}$ 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
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 \( \tilde{\mathcal{K}} \). (c) Visualization of the transformed source and target samples with the domain-invariant kernel \( \tilde{\mathcal{K}} \).

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 $G : \mathcal{H}_1 \rightarrow \mathcal{H}_2$ be a linear operator with finite rank, and let $G^* : \mathcal{H}_2 \rightarrow \mathcal{H}_1$ be its adjoint operator. Then $\text{Im}(G) = \text{Im}(GG^*)$.

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

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

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

Lemma 2. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two separable Hilbert spaces. Let $G : \mathcal{H}_1 \rightarrow \mathcal{H}_2$ be a linear operator with finite rank, and let $G^* : \mathcal{H}_2 \rightarrow \mathcal{H}_1$ 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,

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

(2)

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

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

**Remark 1.** The tensor product $\otimes$ is defined such that $(u \otimes v)w = u\langle v, w \rangle_H, \forall u, v, w \in H$, which is the analogy of the outer product in $\mathbb{R}^n$, i.e., $\vec{u} \otimes \vec{v} = \vec{u}\vec{v}^T, \forall \vec{u}, \vec{v} \in \mathbb{R}^n$. In our paper, these two expressions $u \otimes v$ and $uv^T, \forall u, v \in H$, 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 \{\varphi_k(G^*G)\}_{k=1}^r are eigenvectors of $GG^*$, we need to show that $G\varphi_k(G^*G)$ is nonzero, $\forall k = 1, 2, \ldots, r$. 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 \{\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 $\lambda$ is an positive eigenvalue of $GG^*$, then $\lambda$ 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 $k, l = 1, 2, \ldots, r$,

$$
\langle G\varphi_k(G^*G), G\varphi_l(G^*G) \rangle_{H^1} = \frac{1}{\lambda_k(G^*G)} \langle \varphi_k(G^*G), G\varphi_l(G^*G) \rangle_{H^1} = \frac{1}{\lambda_k(G^*G)} \langle \varphi_k(G^*G), \lambda_k(G^*G)\varphi_l(G^*G) \rangle_{H^1} = 1,
$$

$$
\langle G\varphi_k(G^*G), G\varphi_l(G^*G) \rangle_{H^1} = \lambda_l(G^*G) \langle \varphi_k(G^*G), G\varphi_l(G^*G) \rangle_{H^1} = 0.
$$

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

$$
P_{GG^*} = \sum_{k=1}^r \frac{[G\varphi_k(G^*G)]}{\sqrt{\lambda_k(G^*G)}} \otimes \frac{[G\varphi_k(G^*G)]}{\sqrt{\lambda_k(G^*G)}} = G[\sum_{k=1}^r \lambda_k(G^*G)^{-1} \varphi_k(G^*G) \otimes \varphi_k(G^*G)]G^* = G(G^*)^\dagger G^*.
$$

**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 $\text{Im}(GG^*)$. We can obtain $P_{GG^*}$ immediately.

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

$$
(GG^*)^{\frac{1}{2}} = \sum_{k=1}^r \frac{[\lambda_k(G^*G)^{-\frac{1}{2}} G\varphi_k(G^*G)]}{\sqrt{\lambda_k(G^*G)}} \otimes \frac{[\lambda_k(G^*G)^{-\frac{1}{2}} G\varphi_k(G^*G)]}{\sqrt{\lambda_k(G^*G)}} = G[\sum_{k=1}^r \lambda_k(G^*G)^{-\frac{1}{2}} \varphi_k(G^*G) \otimes \varphi_k(G^*G)]G^* = G(G^*)^\dagger G^*.
$$

**Proof.** Immediately by Lemma 2.

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

$$
(GG^*)^{\frac{1}{2}} = \sum_{k=1}^r \lambda_k(G^*G)^{-\frac{1}{2}} G\varphi_k(G^*G) \otimes G\varphi_k(G^*G) = G[\sum_{k=1}^r \lambda_k(G^*G)^{-\frac{1}{2}} \varphi_k(G^*G) \otimes \varphi_k(G^*G)]G^* = G(G^*)^\dagger G^*.
$$
Proof. Immediately by Lemma 2.

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

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

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 \( G : \mathcal{H}_1 \rightarrow \mathcal{H}_2 \) 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 \( \hat{T}_{WC} = \Sigma_{\ell}^{1/2} (\Sigma_{\ell}^{2/3})^\dagger \).

**Theorem 1.** If \( \text{Im}(\Sigma_{\ell}) \subseteq \text{Im}(\Sigma_{\ell}) \), then \( \hat{T}_{WC} \Sigma_{\ell} \Sigma_{\ell}^{\dagger} = \Sigma_{\ell} \).

**Proof.** Substituting \( \hat{T}_{WC} = \Sigma_{\ell}^{1/2} (\Sigma_{\ell}^{2/3})^\dagger \) into the left part, we have
\[
\hat{T}_{WC} \Sigma_{\ell} \Sigma_{\ell}^{\dagger} = \left[ \Sigma_{\ell}^{1/2} (\Sigma_{\ell}^{2/3})^\dagger \right] \Sigma_{\ell} \left[ \Sigma_{\ell}^{2/3} (\Sigma_{\ell}^{1/2})^\dagger \right]^T = \Sigma_{\ell}^{1/2} (\Sigma_{\ell}^{2/3})^\dagger \Sigma_{\ell} (\Sigma_{\ell}^{1/2})^\dagger = \Sigma_{\ell}^{3/2} P_{\Sigma_{\ell}} \Sigma_{\ell}^{1/2} = \Sigma_{\ell},
\]
where \( P_{\Sigma_{\ell}} \) is the projection matrix onto the image space \( \text{Im}(\Sigma_{\ell}) \), and the last equality holds because \( \text{Im}(\Sigma_{\ell}) \subseteq \text{Im}(\Sigma_{\ell}) \).

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

Given two positive definite covariance matrix \( \Sigma_{\ell} \) and \( \Sigma_{s} \), we have
\[
\Sigma_{\ell}^{1/2} (\Sigma_{\ell}^{2/3} \Sigma_{s}^{1/2} \Sigma_{s}^{1/2})^{-\frac{1}{2}} \Sigma_{s}^{1/2} = \Sigma_{s}^{-\frac{1}{2}} (\Sigma_{s}^{2/3} \Sigma_{\ell}^{1/2} \Sigma_{s}^{1/2})^{-\frac{1}{2}} \Sigma_{s}^{1/2}.
\]
Proof. We write \( \Sigma_t = \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2})^T \), and then substitute it into the right part of (8). We have

\[
\Sigma_t^{-\frac{1}{2}} (\Sigma_t^\frac{1}{2} \Sigma_t \Sigma_t^\frac{1}{2})^2 \Sigma_t^{-\frac{1}{2}} = \Sigma_t^{-\frac{1}{2}} [(\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^2 (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T]^{\frac{1}{2}} \Sigma_t^{-\frac{1}{2}}
\]

(9a)

\[
\Sigma_t^{-\frac{1}{2}} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^2 [(\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})]^{-\frac{1}{2}} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \Sigma_t^{-\frac{1}{2}}
\]

(9b)

\[
\Sigma_t^{-\frac{1}{2}} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^2 (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T - \frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2}
\]

(9c)

\[
\Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} - \Sigma_t^\frac{1}{2})^2
\]

(9d)

where (9b) holds because of Corollary 2.

3.3. Proving Theorem 2

In \( \mathbb{R}^n \), the optimal transport map is \( \hat{T}_{OT} = \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \).

Theorem 2. If \( \text{Ker}(\Sigma_t) \cap \text{Im}(\Sigma_t) = \{0\} \), then we have \( \hat{T}_{OT} \Sigma_t \hat{T}_{OT}^T = \Sigma_t \).

Proof. Let \( \Sigma_t = V_{n \times r} \Lambda_{r \times r} V_{n \times r}^T \), where \( \Lambda \) is a diagonal matrix whose diagonal terms are the \( r \) positive eigenvalues and \( V \) consists of the corresponding eigenvectors.

(I) Claim: rank(\( \Sigma_t^\frac{1}{2} V \)) = \( r \).

Let \( \tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_r \) be the columns of \( V \). It is sufficient to show that \( \Sigma_t^\frac{1}{2} \tilde{v}_1, \Sigma_t^\frac{1}{2} \tilde{v}_2, ..., \Sigma_t^\frac{1}{2} \tilde{v}_r \) are linearly independent.

Suppose there exist \( \lambda_1, \lambda_2, ..., \lambda_r \), such that \( \lambda_1 \Sigma_t^\frac{1}{2} \tilde{v}_1 + \lambda_2 \Sigma_t^\frac{1}{2} \tilde{v}_2 + ... + \lambda_r \Sigma_t^\frac{1}{2} \tilde{v}_r = \tilde{0} \). Then we have that \( \lambda_1 \tilde{v}_1 + \lambda_2 \tilde{v}_2 + ... + \lambda_r \tilde{v}_r \in \text{Ker}(\Sigma_t^\frac{1}{2}) = \text{Ker}(\Sigma_t) \). Since \( \text{Im}(\Sigma_t) = \text{Span}\{\tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_r\} \), immediately, \( \lambda_1 \tilde{v}_1 + \lambda_2 \tilde{v}_2 + ... + \lambda_r \tilde{v}_r \in \text{Im}(\Sigma_t) \). By the condition that \( \text{Ker}(\Sigma_t) \cap \text{Im}(\Sigma_t) = \{0\} \), we have \( \lambda_1 = \lambda_2 = ... = \lambda_r = 0 \). And \( \tilde{v}_1, \tilde{v}_2, ..., \tilde{v}_r \) are linearly independent if \( \lambda_1 = \lambda_2 = ... = \lambda_r = 0 \).

(II) Now we start to prove that \( \hat{T}_{OT} \Sigma_t \hat{T}_{OT}^T = \Sigma_t \). Substituting \( \hat{T}_{OT} = \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \) into the left part, we obtain

\[
\hat{T}_{OT} \Sigma_t \hat{T}_{OT}^T = \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \Sigma_t^\frac{1}{2} = \Sigma_t^\frac{1}{2} P_{ts} \Sigma_t^\frac{1}{2},
\]

(10)

where \( P_{ts} \) is the projection matrix onto \( \text{Im}(\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2}) = \text{Im}[(\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2}) (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T] \). We set \( G = \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \), then we obtain

\[
P_{ts} = \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \Sigma_t^\frac{1}{2}
\]

by Corollary 1. So,

\[
\hat{T}_{OT} \Sigma_t \hat{T}_{OT}^T = \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \Sigma_t^\frac{1}{2} (\Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2})^T \Sigma_t^\frac{1}{2} = \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t^\frac{1}{2} \Sigma_t.
\]

(11)

Substituting \( \Sigma_t = V_{n \times r} A_{r \times r} V_{n \times r}^T \) into the above, we get

\[
\hat{T}_{OT} \Sigma_t \hat{T}_{OT}^T = V \Lambda \Sigma_t^\frac{1}{2} V^T (\Sigma_t^\frac{1}{2} V \Lambda \Sigma_t^\frac{1}{2})^T (\Sigma_t^\frac{1}{2} V \Lambda \Sigma_t^\frac{1}{2})^T
\]

\[
= V \Lambda \Sigma_t^\frac{1}{2} A \Sigma_t^\frac{1}{2} V \Lambda \Sigma_t^\frac{1}{2} V^T
\]

\[
= V \Lambda \Sigma_t^\frac{1}{2} P \Lambda \Sigma_t^\frac{1}{2} V^T,
\]

(12)

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

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

\[ MC_s = (\Phi_X J_{N_s})(\Phi_X J_{N_s})^T + \rho I_{H_C} \]  \hspace{1cm} (13a)
\[ MC_t = (\Phi_Y J_{N_t})(\Phi_Y J_{N_t})^T. \]  \hspace{1cm} (13b)

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

\[ k\hat{T}_{WC} = (MC_t)^{\frac{1}{2}}(MC_s)^{\frac{1}{2}} = \Phi_Y J_{N_t} C_{YY}^{\frac{1}{2}} \{ \sum_k (\Phi_X A J_{N_s} \Phi_X^T + \frac{1}{\sqrt{\rho}} J_{N_t} \Phi_Y^T) \}, \]  \hspace{1cm} (14)

where \( C_{YY} = J_{N_s}^T K_{YY} J_{N_t} \) and \( C_{XX} = J_{N_t}^T K_{YY} J_{N_s} \) are centered kernel matrices, and \( A = \sum_k \frac{1}{\lambda_k} (\frac{1}{\sqrt{\lambda_k + \rho}} - \frac{1}{\sqrt{\rho}}) \bar{v}_k \bar{v}_k^T \), and \( \{ \lambda_k, \bar{v}_k \}_{k=1} \) are positive eigenpairs of \( C_{XX} \).

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

\[ k\hat{T}_{WC} = (\Phi_Y J_{N_t})(\Phi_X J_{N_s})^T + \rho I_{H_C})} \]  \hspace{1cm} (15)

For simplicity, we set \( G = \Phi_X J_{N_s}, \) then by Lemma 3, \( G^* = (\Phi_X J_{N_s})^T \). Let \( \{ \lambda_k \}_{k=1} \) and \( \{ \bar{v}_k \}_{k=1} \) be the positive eigenvalues and the corresponding orthonormal eigenvectors of \( G^*G = (\Phi_X J_{N_s})^T(\Phi_X J_{N_s}) = C_{XX} \). Then by Lemma 2, \( \{ \lambda_k, \bar{v}_k \}_{k=1} \) are the positive eigenpairs of \( GG^* \). Let \( \{ \psi_k \}_{k=r+1}^{\text{dim}(H_C)} \) be a set of orthonormal vectors, such that

\[ \{ \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}}, \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}}, \ldots, \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}}, \psi_{r+1}, \psi_{r+2}, \ldots \} \]  \hspace{1cm} (16)

is a complete orthonormal basis of \( H_C \). Then the identity operator \( I_{H_C} \) can be written as \( I_{H_C} = \sum_{k=1}^{r} \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} \otimes \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} + \sum_{k=r+1}^{\text{dim}(H_C)} \psi_k \otimes \psi_k \). Therefore, we have

\[ \left[ (\Phi_X J_{N_s})(\Phi_X J_{N_s})^T + \rho I_{H_C} \right]^{-\frac{1}{2}} \]  \hspace{1cm} (17a)
\[ = (GG^* + \rho I_{H_C})^{-\frac{1}{2}} \]  \hspace{1cm} (17b)
\[ = \left[ \sum_{k=1}^{r} \lambda_k \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} \otimes \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} + \rho I_{H_C} \right]^{-\frac{1}{2}} \]  \hspace{1cm} (17c)
\[ = \left[ \sum_{k=1}^{r} \sum_{j=1}^{\text{dim}(H_C)} \lambda_k \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} \otimes \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} + \sum_{k=r+1}^{\text{dim}(H_C)} \rho \psi_k \otimes \psi_k \right]^{-\frac{1}{2}} \]  \hspace{1cm} (17d)
\[ = \sum_{k=1}^{r} \frac{1}{\sqrt{\lambda_k + \rho}} \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} \otimes \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} + \sum_{k=r+1}^{\text{dim}(H_C)} \frac{1}{\sqrt{\rho}} \psi_k \otimes \psi_k \]  \hspace{1cm} (17e)
\[ = \sum_{k=1}^{r} \frac{1}{\sqrt{\lambda_k + \rho}} \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} \otimes \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} + \frac{1}{\sqrt{\rho}} (I_{H_C} - \sum_{k=1}^{r} \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}} \otimes \frac{G^* \bar{v}_k}{\sqrt{\lambda_k}}) \]  \hspace{1cm} (17f)
\[ = G \left( \sum_{k=1}^{r} \frac{1}{\lambda_k \sqrt{\lambda_k + \rho}} \bar{v}_k \bar{v}_k^T \right) G^* + \frac{1}{\sqrt{\rho}} [I_{H_C} - G (\sum_{k=1}^{r} \frac{1}{\lambda_k} \bar{v}_k \bar{v}_k^T)] G^* \]  \hspace{1cm} (17g)
\[ = GAG^* + \frac{1}{\sqrt{\rho}} I_{H_C} \]  \hspace{1cm} (17h)
\[ = \Phi_X J_{N_s} A J_{N_s} \Phi_X^T + \frac{1}{\sqrt{\rho}} I_{H_C}. \]  \hspace{1cm} (17i)
Substituting (17i) into (15), we get
\[
k\hat{T}_{WC} = \left[ (\Phi_Y J_{N_i}) (\Phi_Y J_{N_i})^T \right]^{\frac{1}{2}} \left[ (\Phi_X J_{N_i}) (\Phi_X J_{N_i})^T + \rho I_{H\kappa} \right]^{-\frac{1}{2}}
\] (18a)
\[
= \left[ (\Phi_Y J_{N_i}) (\Phi_Y J_{N_i})^T \right]^{\frac{1}{2}} \left( \Phi_X J_{N_i} A J_{N_i}^T \Phi_X^T + \frac{1}{\sqrt{\rho}} I_{H\kappa} \right)
\] (18b)
\[
= \Phi_Y J_{N_i} \left[ J_{N_i}^T, (\Phi_Y J_{N_i}) \right]^{\frac{1}{2}} \left( \Phi_Y J_{N_i} A J_{N_i}^T \Phi_X^T + \frac{1}{\sqrt{\rho}} I_{H\kappa} \right)
\] (18c)
\[
= \Phi_Y J_{N_i} C_{YY}^{-\frac{1}{2}} \left[ C_{YX} A J_{N_i} \Phi_X^T + \frac{1}{\sqrt{\rho}} J_{N_i} \Phi_Y^T \right],
\] (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
\[
EC_s = (\Phi_X W_X)(\Phi_X W_X)^T + \rho I_{H\kappa}
\] (19a)
\[
EC_t = (\Phi_Y W_Y)(\Phi_Y W_Y)^T.
\] (19b)

**Proposition 2.** With the computationally efficient estimators \( (19) \), the kernel optimal transport map is given by
\[
k\hat{T}_{OT} = (EC_t)^{\frac{1}{2}} \left[ (EC_t)^{\frac{1}{2}} (EC_s)/(EC_t)^{\frac{1}{2}} \right]^{\frac{1}{2}} (EC_t)^{\frac{1}{2}} = \Phi_Y W_Y \left[ C_{YX}^w C_{XY}^w + \rho \Lambda_Y - \rho I_d \right]^{\frac{1}{2}} \Phi_Y W_Y^T \Phi_Y^T,
\] (20)
where \( C_{YX}^w = W_Y^T K_{YX} W_X \) and \( C_{XY}^w = (C_{YX})^T \), and \( \Lambda_Y \) is the diagonal matrix storing the top \( d \) eigenvalues of \( C_{YX} \).

**Proof.** We substitute the computationally efficient estimators \( (19) \) into the expression of \( k\hat{T}_{OT} \), and then we have,
\[
k\hat{T}_{OT} = (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T) \right]^{\frac{1}{2}} \left( \Phi_Y W_Y (\Phi_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right)^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (21)
We observe that \( EC_s = \Phi_X W_X (\Phi_X W_X)^T + \rho I_{H\kappa} \) is strictly positive definite, which implies that there exists an operator \( A \) (e.g., \( A = (EC_s)^{\frac{1}{2}} \)), such that \( A A^T = EC_s \). Then \( k\hat{T}_{OT} \) can be written as
\[
k\hat{T}_{OT} = (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T) \right]^{\frac{1}{2}} \left( \Phi_Y W_Y (\Phi_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right)^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22a)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} A \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22b)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22c)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} A \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22d)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T) A \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22e)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T) \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22f)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T) \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22g)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T) \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22h)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T) \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22i)
\[
= (\Phi_Y W_Y W_Y^T \Phi_Y^T) \left[ (\Phi_Y W_Y W_Y^T \Phi_Y^T)^{\frac{1}{2}} \right]^{\frac{1}{2}} \left( \Phi_Y W_Y W_Y^T \Phi_Y^T \right)^{\frac{1}{2}}
\] (22j)
where (22c) and (22g) hold because of Corollary 3. Next we prove that \( W_{s}^{T}K_{YY}W_{s} = \Lambda_{Y} - \rho I_{d} \):

\[
W_{s}^{T}K_{YY}W_{s} = [J_{N_{t}}V_{Y}(I_{d} - \rho \Lambda_{Y}^{-1})^{\frac{1}{2}}]^{T}K_{YY}[J_{N_{t}}V_{Y}(I_{d} - \rho \Lambda_{Y}^{-1})^{\frac{1}{2}}] \\
= (I_{d} - \rho \Lambda_{Y}^{-1})^{\frac{1}{2}}V_{Y}^{T}C_{YY}V_{Y}^{T}(I_{d} - \rho \Lambda_{Y}^{-1})^{\frac{1}{2}} \\
= (I_{d} - \rho \Lambda_{Y}^{-1})^{\frac{1}{2}}\Lambda_{Y}(I_{d} - \rho \Lambda_{Y}^{-1})^{\frac{1}{2}} \\
= \Lambda_{Y} - \rho I_{d}.
\] (23)

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

3.6. Derivations of the domain-invariant kernel expressions

The domain-invariant kernel \( \Delta K \) is given by

\[
\Delta K = \begin{bmatrix} \Delta K_{ss} & \Delta K_{ts} \\ \Delta K_{ts}^{T} & \Delta K_{tt} \end{bmatrix} = \begin{bmatrix} \Psi_{s \rightarrow t}^{T}\Psi_{s \rightarrow t} & \Psi_{s \rightarrow t}^{T}\Psi_{t} \\ \Psi_{t}^{T}\Psi_{s \rightarrow t} & \Psi_{t}^{T}\Psi_{t} \end{bmatrix},
\] (24)

where the symbol \( \Delta \) represents the way of “moving” the source samples, i.e., \( \Delta = WC \) or \( OT \), and \( \Psi_{s \rightarrow t} \) denotes the transported source samples, i.e., \( \Psi_{s \rightarrow t} = kT_{\Delta}(\Psi_{s}) \), and \( \Psi_{s} \) and \( \Psi_{t} \) denote the centered source samples and target samples, respectively.

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

\[
\Psi_{s \rightarrow t} = kT_{WC}(\Psi_{s}) = \sqrt{N_{s}}\Phi_{Y}J_{N_{t}}C_{Y}^{\frac{1}{2}}B \\
WC\Delta K_{ss} = N_{s}B^{T}B \\
WC\Delta K_{ts} = \sqrt{N_{s}N_{t}}C_{Y}^{\frac{1}{2}}B = \sqrt{N_{s}N_{t}}U_{Y}\Lambda_{Y}^{\frac{1}{2}}U_{Y}^{T}B,
\] (25)

where \( B = C_{Y}^{\frac{1}{2}}(C_{XX} + \rho I_{N_{t}})^{-\frac{1}{2}} \), and \( (U_{Y}, \Lambda_{Y}^{\frac{1}{2}}) \) stores the top \( d \) eigenpairs of \( C_{YY} \). 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

\[
\Psi_{s \rightarrow t} = kT_{WC}(\Psi_{s}) \\
= \Phi_{Y}J_{N_{t}}C_{Y}^{\frac{1}{2}}[C_{Y}XAJ_{N_{t}}\Phi_{X}^{T} + \frac{1}{\sqrt{\rho}}J_{N_{t}}\Phi_{X}^{T}](\sqrt{N_{s}}\Phi_{X}J_{N_{t}}) \\
= \sqrt{N_{s}}\Phi_{Y}J_{N_{t}}C_{Y}^{\frac{1}{2}}C_{Y}X(AC_{XX} + \frac{1}{\sqrt{\rho}}I_{N_{t}}).
\] (26a, 26b, 26c)

Now we consider the term \( AC_{XX} + \frac{1}{\sqrt{\rho}}I_{N_{t}} \). Recall that \{\vec{v}_{1}, \vec{v}_{2}, ..., \vec{v}_{r}\} are \( C_{XX} \)'s eigenvectors, the corresponding eigenvalues of which are positive. Let \{\vec{u}_{r+1}, \vec{u}_{r+2}, ..., \vec{u}_{N_{t}}\} be a set of orthonormal vectors, such that \{\vec{v}_{1}, \vec{v}_{2}, ..., \vec{v}_{r}, \vec{u}_{r+1}, \vec{u}_{r+2}, \}
..., \{ \tilde{u}_{N_r} \} \) is an orthonormal system for \( \mathbb{R}^{N_r} \). Then, we have

\[
AC_{XX} + \frac{1}{\sqrt{\rho}} I_{N_r} = \sum_{k=1}^{r} \frac{1}{\sqrt{\lambda_k + \rho}} \tilde{v}_k \tilde{v}_k^T C_{XX} + \frac{1}{\sqrt{\rho}} I_{N_r}, \tag{27a}
\]

\[
= \sum_{k=1}^{r} \frac{1}{\sqrt{\lambda_k + \rho}} \tilde{v}_k (C_{XX} \tilde{v}_k)^T + \frac{1}{\sqrt{\rho}} I_{N_r}, \tag{27b}
\]

\[
= \sum_{k=1}^{r} \frac{1}{\sqrt{\lambda_k + \rho}} \tilde{v}_k \tilde{v}_k^T + \frac{1}{\sqrt{\rho}} \sum_{k=1}^{r} \tilde{u}_k \tilde{u}_k^T + \frac{1}{\sqrt{\rho}} \sum_{k=r+1}^{N_r} \tilde{u}_k \tilde{u}_k^T, \tag{27c}
\]

\[
\sum_{k=1}^{r} \frac{1}{\sqrt{\lambda_k + \rho}} \tilde{v}_k \tilde{v}_k^T + \frac{1}{\sqrt{\rho}} \sum_{k=r+1}^{N_r} \tilde{u}_k \tilde{u}_k^T \tag{27d}
\]

\[
\equiv (C_{XX} + \rho I_{N_r})^{-\frac{1}{2}}. \tag{27e}
\]

Write \( B = C_{YY}(C_{XX} + \rho I_{N_r})^{-\frac{1}{2}} \), then \( \Psi_{s \rightarrow t} = k_T \psi_{WC} = \sqrt{N_s} \Phi_Y J_{N_r} C_{YY}^{\frac{1}{2}} B \).

(II)

\[
WC \vec{K}_{ss} = \Psi_{s \rightarrow t}^{T} \Psi_{s \rightarrow t} = (\sqrt{N_s} \Phi_Y J_{N_r} C_{YY}^{\frac{1}{2}} B)^T (\sqrt{N_s} \Phi_Y J_{N_r} C_{YY}^{\frac{1}{2}} B) \tag{28}
\]

\[
= N_s B^T C_{YY}^{\frac{1}{2}} J_{N_r}^{\dagger} \Phi_Y J_{N_r} C_{YY}^{\frac{1}{2}} B = N_s B^T P_{YY} B = N_s B^T B,
\]

where \( P_{YY} \) is the projection matrix onto \( \text{Im}(C_{YY}) = \text{Im}(J_{N_r}^{\dagger} \Phi_Y) \), and the last equality holds because \( \text{Im}(B) \subseteq \text{Im}(C_{YY}) \subseteq \text{Im}(J_{N_r}^{\dagger} \Phi_Y) \).

(III)

\[
WC \vec{K}_{ts} = \Psi_{t}^{T} \Psi_{s \rightarrow t} = (\sqrt{N_t} \Phi_Y J_{N_t} C_{YY}^{\frac{1}{2}} B) = \sqrt{N_t} N_s C_{YY}^{\frac{1}{2}} B. \tag{29}
\]

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

\[
\Psi_{s \rightarrow t} = k_T \psi_{OT} = \sqrt{N_s} \Phi_Y W_Y D \tag{30}
\]

\[
OT \vec{K}_{ss} = N_s D^T (\Lambda_Y - \rho I_d) D
\]

\[
OT \vec{K}_{ts} = \sqrt{N_s N_t J_{N_r} K_{YY} W_Y D},
\]

where \( D = [C_{YY}^{w} C_{XX}^{w} + \rho (\Lambda_Y - \rho I_d)]^{\frac{1}{2}} W^T_Y K_{YY} J_{N_r} \).

Proof.

(I) The transported samples are

\[
\Psi_{s \rightarrow t} = k_T \psi_{OT} = \Phi_Y W_Y [C_{YY}^{w} C_{XX}^{w} + \rho (\Lambda_Y - \rho I_d)]^{\frac{1}{2}} W^T_Y (\sqrt{N_s} \Phi_X J_{N_r}) \tag{31}
\]

\[
= \sqrt{N_s} \Phi_Y W_Y [C_{YY}^{w} C_{XX}^{w} + \rho (\Lambda_Y - \rho I_d)]^{\frac{1}{2}} W^T_Y K_{YY} J_{N_r} = \sqrt{N_s} \Phi_Y W_Y D.
\]

(II)

\[
OT \vec{K}_{ss} = \Psi_{s \rightarrow t}^{T} \Psi_{s \rightarrow t} = (\sqrt{N_s} \Phi_Y W_Y D)^T (\sqrt{N_s} \Phi_Y W_Y D) = N_s D^T W^T_Y K_{YY} W_Y D = N_s D^T (\Lambda_Y - \rho I_d) D, \tag{32}
\]

where the last equality holds because of (23).

(III)

\[
OT \vec{K}_{ts} = \Psi_{t}^{T} \Psi_{s \rightarrow t} = (\sqrt{N_t} \Phi_Y J_{N_t})^T (\sqrt{N_s} \Phi_Y W_Y D) = \sqrt{N_t} N_s J_{N_r} K_{YY} W_Y D. \tag{33}
\]
References


