# Introduction of Two Domain Adaptation Methods
## Personal Reading Notes

*Jianglin Lu*

*Department of ECE*
*College of Engineering*
*Northeastern University*
*360 Huntington Avenue, Boston, MA 02115, USA*
*https: // jianglin954. github. io/*
*jianglinlu@outlook.com*

## Outline

**1** Preliminaries

**2** Domain Adaptation under Target and Conditional Shift

**3** Domain Adaptation with Conditional Transferable Components

# Outline

**1** Preliminaries

**2** Domain Adaptation under Target and Conditional Shift

**3** Domain Adaptation with Conditional Transferable Components

# Preliminaries

## Kernel

Given a non-empty set $\mathcal{X}$, $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is a *kernel* if:

- $\kappa$ is symmetric: $\kappa(x, y) = \kappa(y, x)$.
- $\kappa$ is positive semi-definite, i.e., $\forall x_1, x_2, ..., x_n \in \mathcal{X}$, the "Gram Matrix" $K$ defined by $K_{ij} = \kappa(x_i, x_j)$ is positive semi-definite.

Properties:

- $\kappa(x, x) \geq 0$.
- $\kappa(u, v) \leq \sqrt{\kappa(u, u) \cdot \kappa(v, v)}$ (Cauchy-Schwarz inequality).

* A matrix $M \in \mathbb{R}^{n \times n}$ is positive semi-definite if $\forall a \in \mathbb{R}^n$, $a'Ma \geq 0$.

# Preliminaries

**Kernel methods[1]:**

Owe their name to the use of kernel functions, which enable them to operate in a high-dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This operation is often computationally cheaper than the explicit computation of the coordinates.

Kernel methods can (and often do) use infinitely many features. This can be achieved as long as our learning algorithms are defined in terms of dot products between the features, where these dot products can be computed in closed form. The term kernel simply refers to a dot product between (possibly infinitely many) features.

---

[1] Wikipedia, https://en.wikipedia.org/wiki/Kernel_method.

## Preliminaries

### Inner product

Let $\mathcal{F}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle_{\mathcal{F}} : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{R}$ is an *inner product* on $\mathcal{F}$ if:

- Linear: $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{F}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{F}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{F}}$.
- Symmetric: $\langle f, g \rangle_{\mathcal{F}} = \langle g, f \rangle_{\mathcal{F}}$.
- $\langle f, f \rangle_{\mathcal{F}} \geq 0$ and $\langle f, f \rangle_{\mathcal{F}} = 0$ if and only if $f = 0$.

$*$ Vector space with an inner product is said to be an inner product space.

$*$ $\langle f, g \rangle_{\mathcal{F}} = 0$, $\forall f \in \mathcal{F}$ if and only if $g = 0$.

$*$ $\langle f, \alpha_1 g_1 + \alpha_2 g_2 \rangle_{\mathcal{F}} = \alpha_1 \langle f, g_1 \rangle_{\mathcal{F}} + \alpha_2 \langle f, g_2 \rangle_{\mathcal{F}}$.

# Preliminaries

## Norm

*Norm* induced by the inner product: $||x||_{\mathcal{H}} := \sqrt{\langle x, x \rangle_{\mathcal{H}}}$.

## Normed space

A *normed space* is a linear (vector) space $\mathcal{H}$ in which a norm is defined.
A nonnegative function $|| \cdot ||$ is a norm *iff* $\forall f, g \in \mathcal{H}$ and $\alpha \in \mathbb{R}$:

- $||f|| \geq 0$ and $||f|| = 0$ *iff* $f = 0$.
- $||f + g|| \leq ||f|| + ||g||$.
- $||\alpha f|| = |\alpha| \ ||f||$.

## Preliminaries

### Kernel function

A *kernel* is a *function* $\kappa$ that for all $x$, $z \in \mathcal{X}$ satisfies:

$$\kappa(x, z) = \langle \phi(x), \phi(z) \rangle, \tag{1}$$

where $\phi$ is a mapping from $\mathcal{X}$ to an (inner product) feature space $\mathcal{H}$:

$$\phi :\longmapsto \phi(x) \in \mathcal{H}. \tag{2}$$

∗ Kernel functions[2] make possible the use of feature spaces with an exponential or even infinite number of dimensions.

∗ Compute the inner product between the projections of two points into the feature space without explicitly evaluating their coordinates.

∗ The feature space is not uniquely determined by the kernel function.

---

[2]Taylor et al. Kernel Methods for Pattern Analysis, Cambridge University Press, 2004

## Preliminaries

### Hilbert space

A *Hilbert Space* $\mathcal{H}$ is an inner product space that is *complete* and *separable* with respect to the norm defined by the inner product:

$$||f||_{\mathcal{H}} = \sqrt{\langle f, f \rangle_{\mathcal{H}}}, \quad \forall f \in \mathcal{H} \tag{3}$$

$*$ Hilbert spaces generalize the finite Euclidean spaces $\mathbb{R}^d$, and are generally *infinite dimensional*.

$*$ Separability implies that Hilbert spaces have *countable orthonormal bases*.

$*$ $\mathcal{H}$ is a complete space if every Cauchy sequence in $\mathcal{H}$ is convergent. Informally, a space is complete if every (infinite) sequence of its elements that approaches a particular value has this value as it's limit and this limit is in the space itself.

# Preliminaries

### Cauchy sequence

A *Cauchy sequence* is a sequence $(x_i)_{i=0}^{\infty}$ such that for every real number $\epsilon > 0$ we can find a natural number $N$ such that $d(x_n, x_m) < \epsilon$ whenever $n$, $m > N$. Here, $d$ is a distance metric on $\mathcal{H}$.

$*$ This basically says that we can take an arbitrarily small value for $\epsilon$ and are guaranteed that after some point ($N$), all later values of $x$ are no further apart than $\epsilon$.

## Preliminaries

### Convergent sequence

A sequence is *convergent* in $X$ if there is a point $x \in X$ such that for every real number $\epsilon > 0$ we can find a natural number $N$ such that $d(x, x_n) < \epsilon$ for all $n > N$.

$*$ This says that for any convergent sequence, we can find some value $x$ that is *in the original space* that is arbitrarily close to $x_n$ for all $n$ after a certain point.

# Preliminaries

## Reproducing property

The evaluation of $f$ at $x$ can be written as an *inner product in feature space:*
$$\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \langle f, \kappa(\cdot, x) \rangle_{\mathcal{H}} = f(x) \tag{4}$$

## Preliminaries

### Reproducing kernel

Let $\mathcal{H}$ be a Hilbert space of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$. A function $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if it satisfies:

- $\forall x \in \mathcal{X}, \ \kappa(\cdot, x) \in \mathcal{H}$.
  (informally, the feature map of every point is in the feature space.)
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}, \ \langle f(\cdot), \kappa(\cdot, x) \rangle_{\mathcal{H}} = f(x)$
  (i.e., the reproducing property).

In particular, for any $x, y \in \mathcal{X}$, $\kappa(x, y) = \langle \kappa(\cdot, x), \kappa(\cdot, y) \rangle_{\mathcal{H}}$.

* If it exists, reproducing kernel is unique.
* Reproducing kernels are positive definite.
* We can write $\phi(x) = \kappa(\cdot, x)$ without ambiguity.

## Preliminaries

### Evaluation functional

Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \longrightarrow \mathbb{R}$, defined on a non-empty set $\mathcal{X}$. For a fixed $x \in \mathcal{X}$, map $\delta_x : \mathcal{H} \longrightarrow \mathbb{R}$, $\delta_x : f \longmapsto f(x)$ is called the *evaluation functional* at $x$ (i.e., $\delta_x f = f(x)$).

∗ Evaluation functionals are always linear: For $f$, $g \in \mathcal{H}$ and $\alpha$, $\beta \in \mathbb{R}$, $\delta_x(\alpha f + \beta g) = (\alpha f + \beta g)(x) = \alpha f(x) + \beta g(x) = \alpha \delta_x(f) + \beta \delta_x(g)$.

# Preliminaries

## Reproducing Kernel Hilbert Space

A Hilbert space $\mathcal{H}$ of functions $f : \mathcal{X} \longrightarrow \mathbb{R}$, defined on a non-empty set $\mathcal{X}$ is said to be a *Reproducing Kernel Hilbert Space (RKHS)* if $\delta_x$ is bounded (continuous) $\forall x \in \mathcal{X}$: there exists a corresponding $\lambda_x \geq 0$ such that $\forall f \in \mathcal{H}$, $|f(x)| = |\delta_x f| \leq \lambda_x ||f||_{\mathcal{H}}$.

$*$ If two functions converge in RKHS norm, then they converge at every point, i.e., if $\lim_{n \to \infty} ||f_n - f||_{\mathcal{H}} = 0$, then $\lim_{n \to \infty} f_n(x) = f(x)$, $\forall x \in \mathcal{X}$.

$*$ $\mathcal{H}$ is a reproducing kernel Hilbert space *iff* $\mathcal{H}$ has a reproducing kernel.

$*$ Given a RKHS $\mathcal{H}$, we may define a unique reproducing kernel associated with $\mathcal{H}$, which is a positive definite function.

## Preliminaries

### Reproducing Kernel Hilbert Space

Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert space of functions defined on $\mathcal{H}$. Then $\mathcal{H}$ is a *Reproducing Kernel Hilbert Space (RKHS)*, if there exists a bilinear form $\kappa : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ such that

- $\phi(x) = \kappa(x, \cdot)$, $\phi(x) : \mathcal{X} \longrightarrow \mathcal{H}$ is the feature map.
- $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$, $\langle f(\cdot), \kappa(\cdot, x) \rangle_{\mathcal{H}} = f(x)$,

we call these the reproducing property of $\kappa$.

$*$ We denote the RKHS $\mathcal{H}$ with reproducing kernel $\kappa$ interchangeably by $\mathcal{H}_\kappa = \mathcal{H}_\kappa(\mathcal{X})$. The correspondence of $\kappa$ and $\mathcal{H}_\kappa$ is one-to-one.

# Preliminaries

### Riesz representation

In a Hilbert space $\mathcal{H}$, all bounded linear functionals are of the form $\langle \cdot, g \rangle_{\mathcal{H}}$, for some $g \in \mathcal{H}$.

## Preliminaries

### Moore-Aronszajn Theorem

Let $\mathcal{X}$ be a metric space, and $k : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ be a positive definite function, there exists a unique Hilbert space $(\mathcal{H}_\kappa, \langle \cdot, \cdot \rangle_{\mathcal{H}_\kappa})$ of functions on $\mathcal{X}$ satisfying the followings:

- $\phi(x) = \kappa(x, \cdot) \in \mathcal{H}$, $\forall x \in \mathcal{X}$.
- $Span\{\phi(x) : x \in \mathcal{X}\}$ is dense in $\mathcal{H}$,
- $\langle f(\cdot), \phi(x) \rangle_{\mathcal{H}} = f(x)$, $\forall x \in \mathcal{X}, \forall f \in \mathcal{H}$.

$*$ $\mathcal{H}$ is the unique RKHS with reproducing kernel $\kappa$ (denoted by $\mathcal{H}_\kappa$).

$*$ Every RKHS has a unique positive definite kernel.

$*$ Feature map is *not unique*, only kernel is unique.

# Preliminaries

Remark #1: Given any feature map $\phi(\cdot) : \mathcal{X} \longrightarrow \mathcal{H}$, we can define a positive definite function $\kappa(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}_\kappa}$ and a corresponding RKHS.

Remark #2: For any RKHS with kernel $\kappa$ (which is unique), we may define the feature map to be $\phi(x) = \kappa(x, \cdot)$. This means that for data $\{x_n\}_{n \in \mathbb{N}}$, $\kappa(x, x')$ can be computed directly without the evaluation of $\phi(x)$, $\phi(x')$ and the inner product between them. This is known as the kernel method in ML, which is really useful practically as many feature maps are of infinite dimensions (e.g., Gaussian).

## Preliminaries

The mean, variance and covariance have their counterparts in the feature space. In infinite dimensional RKHSs, they are defined as **mean element, covariance operator, and cross-covariance operator** respectively.

### Mean element

Assume $\mathcal{X}$ is a metric spaces, and $X$ is a random variable supported on $\mathcal{X}$. The *mean element* in $\mathcal{H}_\kappa$ is defined as $\mu_X = \mathbb{E}_X[\kappa(X, \cdot)] \in \mathcal{H}_\kappa$ such that for any $f \in \mathcal{H}_\kappa$:

$$\langle f, \mu_X \rangle_{\mathcal{H}_\kappa} = \mathbb{E}_X[f(X)] \tag{5}$$

$*$ Existence of mean element: if $\kappa$ is uniformly bounded (i.e., $\sup_{x \in \mathcal{X}} \kappa(x, x) < \infty$), $\mu_X \in \mathcal{H}_\kappa$.

# Preliminaries

## Cross-covariance operator

Assume $\mathcal{Y}$ is another metric spaces, and $Y$ is a random variable supported on $\mathcal{Y}$. Let $\mathcal{G}_h$ be another RKHS equipped with reproducing kernel $h$. Let $\mu_X$, $\mu_Y$ be the two mean elements in $\mathcal{H}_\kappa$ and $\mathcal{G}_h$, respectively. Then the *cross-covariance operator* is defined as:

$$C_{XY} = \mathbb{E}_{XY}\left[(\kappa(X, \cdot) - \mu_X) \otimes (h(Y, \cdot) - \mu_Y)\right], \tag{6}$$

where for any $f \in \mathcal{H}_\kappa$, $g \in \mathcal{G}_h$,

$$\begin{aligned}
\langle f, C_{XY}g \rangle_{\mathcal{H}_\kappa} &= \mathbb{E}_{XY}\left[\langle \kappa(X, \cdot) - \mu_X, f \rangle_{\mathcal{H}_\kappa} \langle h(Y, \cdot) - \mu_Y, g \rangle_{\mathcal{G}_h}\right], \\
&= \mathbb{E}_{XY}\left[(f(X) - \mathbb{E}_X[f])(g(Y) - \mathbb{E}_Y[g])\right].
\end{aligned} \tag{7}$$

Therefore, the *(auto-)covariance operator* is as followed

$$\Sigma_X = C_{XX} = \mathbb{E}_{XX}\left[(\kappa(X, \cdot) - \mu_X) \otimes (\kappa(X, \cdot) - \mu_X)\right], \tag{8}$$

## Preliminaries

### Universal kernel

A continuous kernel $\kappa$ on a compact metric space $(\mathcal{X}, d)$ is called *universal* if the space of all functions induced by $\kappa$ is dense in $C(\mathcal{X})$, i.e. for every function $f \in C(\mathcal{X})$ and every $\varepsilon > 0$ there exists a function $g$ induced by $\kappa$ with

$$||f - g||_\infty \leq \varepsilon \qquad (9)$$

∗ A metric $d$ on $\mathcal{X}$ is a function $d : \mathcal{X} \times \mathcal{X} \longrightarrow [0, \infty)$ such that for all $x$, $y$, $z \in \mathcal{X}$ we have $d(x, y) = d(y, x)$ and $d(x, z) \leq d(x, z) + d(z, y)$ as well as $d(x, y) = 0$ if and only if $x = y$.

∗ $C(\mathcal{X})$ means the space of all continuous functions $f : X \longrightarrow \mathbb{R}$ on the compact metric space $(\mathcal{X}, d)$ endowed with the usual supremum norm $||f||_\infty := \sup_{x \in \mathcal{X}} |f(x)|$.

## Preliminaries

### Characteristic kernel

Let $(\Omega, \mathcal{B})$ be a measurable space, let $(\mathcal{H}, \kappa)$ be an RKHS over $\Omega$ with the kernel $\kappa$ measurable and bounded, and let $\xi$ be the set of all probability measures on $(\Omega, \mathcal{B})$. The RKHS $\mathcal{H}$ is called *characteristic* (with respect to $\mathcal{B}$) if the map

$$\xi \ni P \mapsto m_P = \mathbb{E}_{X \sim P}[\kappa(\cdot, X)] \in \mathcal{H} \tag{10}$$

is one-to-one (injective), where $m_P$ is the mean element of the random variable with law $P$.

$*$ We also call a positive definite kernel $\kappa$ characteristic if the associated RKHS is characteristic.

$*$ In other words, a characteristic kernel induces an RKHS that is sufficiently rich in the sense that probability measures have unique images.

# Preliminaries

## Isomorphisms

Given two fields $\mathbb{F}$ and $\mathbb{G}$, we say that $\phi$ is an *isomorphism* between $\mathbb{F}$ and $\mathbb{G}$ if $\phi$ is a function from $F \longrightarrow G$ and $\phi$ obeys certain properties:

- *Injective* (one to one): $\forall f, f' \in F$, $\phi(f) = \phi(f')$ implies that $f = f'$ (i.e., there is at most one element in $F$ which maps to a single element in $G$).

- *Surjective* (onto): $\forall g \in G$, there exists $f \in F$ such that $\phi(f) = g$ (i.e., there is at least one element in $F$ which maps to a single element in $G$). The combination of these first two properties states that $\phi$ is a *bijection*.

- *Preservation*: basically, $\phi$ preserves operations. That is, for example, $\phi(a + b) = \phi(a) + \phi(b)$ and $\phi(ab) = \phi(a)\phi(b)$.

# Hilbert Space Embedding for Distributions

## Kernel Mean Embedding

Suppose that $X = \{x_1, ..., x_m\}$ is drawn independently and identically distributed from $P_X$, we make the following two mappings:

$$\mu[P_X] := \mathbb{E}_X[\kappa(x, \cdot)], \quad \mu[X] := \frac{1}{m} \sum_{i=1}^{m} \kappa(x_i, \cdot) \tag{11}$$

By the reproducing property of $\mathcal{H}$, we have:

$$\langle \mu[P_X], f \rangle = \langle \mathbb{E}_X[\kappa(x, \cdot)], f \rangle = \mathbb{E}_X[\langle \kappa(x, \cdot), f \rangle] = \mathbb{E}_X[f(x)]$$

$$\langle \mu[X], f \rangle = \frac{1}{m} \sum_{i=1}^{m} f(x_i) \tag{12}$$

$*$ If $\mathbb{E}[\kappa(x, x)] < \infty$, $\mu[P_X]$ is an element of the Hilbert space (i.e., probability distributions can be represented as elements in an RKHS).

# Hilbert Space Embedding for Distributions

## Kernel Mean Embedding

*Theorem #1*: If the kernel $\kappa$ is universal, then the mean map $\mu : P_X \longrightarrow \mu[P_X]$ is injective.

*Theorem #2*: Assume that $||f||_\infty \leq R$ for all $f \in \mathcal{H}$ with $||f||_\mathcal{H} \leq 1$. Then with probability at least $1 - \delta$, $||\mu[P_X] - \mu[X]|| \leq 2R_m(\mathcal{H}, P_X) + R\sqrt{-m^{-1}\log(\delta)}$.

$*$ Theorem #1 means that $\mu[P_X]$ can be used to define the distances between distributions $P_X$ and $P_Y$, simply by letting $D(P_X, P_Y) := ||\mu[P_X] - \mu[P_Y]||$.

$*$ Theorem #2 means that we do not need to have access to actual distributions in order to compute $D(P_X, P_Y)$ approximately—as long as $R_m(\mathcal{H}, P_X) = \mathcal{O}(m^{-\frac{1}{2}})$, a finite sample from the distributions will yield error of $\mathcal{O}(m^{-\frac{1}{2}})$.

# Outline

# Background

### Covariate shift / Sample selection bias

When the distributions on training and test sets do not match, we are facing *sample selection bias* or *covariate shift* (i.e., $P_X^{tr} \neq P_X^{te}$, while $P_{Y|X}^{tr} = P_{Y|X}^{te}$).

## Sample Reweighting[1]

In general, a learning method minimizes the expected risk:

$$R[P^{tr}, \theta, l(x, y; \theta)] = \mathbb{E}_{(X,Y) \sim P^{tr}}[l(x, y; \theta)] \tag{13}$$

However, we would like to minimize $R[P^{te}, \theta, l(x, y; \theta)]$ as we wish to generalize to test samples drawn from $P^{te}$. From the view of *importance sampling*, we have

$$\begin{aligned}
R[P^{te}, \theta, l(x, y; \theta)] &= \mathbb{E}_{(X,Y) \sim P^{te}}[l(x, y; \theta)] \\
&= \mathbb{E}_{(X,Y) \sim P^{tr}}\left[\frac{P_{XY}^{te}}{P_{XY}^{tr}} \cdot l(x, y; \theta)\right] \\
&= R[P^{tr}, \theta, \beta(x, y) \cdot l(x, y; \theta)]
\end{aligned} \tag{14}$$

where $\beta(x, y) \triangleq P_{XY}^{te} / P_{XY}^{tr}$.

---

[1] *Huang et al. Correcting Sample Selection Bias by Unlabeled Data, NIPS 2006.*

# Motivation of Zhang et al[2]

We address the situation where both the marginal distribution $P_X$ and the conditional distribution $P_{Y|X}$ may change across the domains.

## Three possible scenarios

- **Target Shift (TarS):** the marginal distribution $P_Y$ changes, while the conditional $P_{X|Y}$ stays the same.

- **Conditional Shift (ConS):** the marginal distribution $P_Y$ is fixed, while the conditional $P_{X|Y}$ changes. A practical case where $P_{X|Y}$ changes under location-scale (LS) transformations on $X$.

- **Generalized Target Shift (GeTarS):** the marginal distribution $P_Y$ changes, and the conditional $P_{X|Y}$ changes. A practical case is LS-GeTarS where $P_{X|Y}$ changes under LS transformations.

---

[2] *Zhang et al. Domain Adaptation under Target and Conditional Shift, ICML 2013.*

## Importance Reweighting

Assume the support of $P_{XY}^{te}$ is contained by that of $P_{XY}^{tr}$. The expected loss on test data is:

$$
\begin{aligned}
R[P^{te}, \theta, l(x, y; \theta)] &= \mathbb{E}_{(X,Y) \sim P^{te}}[l(x, y; \theta)] \\
&= \int P_{XY}^{tr} \cdot \frac{P_{XY}^{te}}{P_{XY}^{tr}} \cdot l(x, y; \theta) dx dy \\
&= \mathbb{E}_{(X,Y) \sim P^{tr}} \left[ \frac{P_Y^{te}}{P_Y^{tr}} \cdot \frac{P_{X|Y}^{te}}{P_{X|Y}^{tr}} \cdot l(x, y; \theta) \right] \\
&= \mathbb{E}_{(X,Y) \sim P^{tr}} [\beta^*(y) \cdot \gamma^*(x, y) \cdot l(x, y; \theta)]
\end{aligned}
\tag{15}
$$

where $\beta^*(y) \triangleq P_Y^{te}/P_Y^{tr}$ and $\gamma^*(x, y) \triangleq P_{X|Y}^{te}/P_{X|Y}^{tr}$.

$*$ The *support* (or sample space) of a random variable is defined as the set of numbers that are possible values of the random variable.

$*$ Here, we factorize $P_{XY}$ as $P_Y P_{X|Y}$ instead of $P_X P_{Y|X}$ because it provides a more convenient way to handle the change in $P_{XY}$.

## Importance Reweighting

In practice, we minimize the empirical loss:

$$\hat{R} = \frac{1}{m} \sum_{i=1}^{m} \beta^*(y_i^{tr}) \gamma^*(x_i^{tr}, y_i^{tr}) l(x_i^{tr}, y_i^{tr}; \theta) \tag{16}$$

if $\beta^*(y_i^{tr}) \gamma^*(x_i^{tr}, y_i^{tr})$ are given.

## Sample Transformation and Reweighting

Drawback of sample reweighting: in the case where both $P_Y$ and $P_{X|Y}$ change, the application of sample reweighting scheme is rather limited.

Solution: find the transformation $\mathcal{T}$ such that the conditional distribution of $X^{new} = \mathcal{T}(X^{tr}, Y^{tr})$ satisfies $P_{X|Y}^{new} = P_{X|Y}^{te}$.

$$R[P^{te}, \theta, l(x, y; \theta)] = \mathbb{E}_{(X,Y) \sim P^{te}}[l(x, y; \theta)]$$
$$= \int P_Y^{tr} \cdot \beta^*(y) \cdot P_{X|Y}^{te} \cdot l(x, y; \theta) dx dy \qquad (17)$$
$$= \mathbb{E}_{(X,Y) \sim P_Y^{tr} P_{X|Y}^{new}}[\beta^*(y) \cdot l(x, y; \theta)]$$

This empirical loss can be calculated on the transformed training points $x^{new}, y^{tr}$ with wights $\beta^*$:

$$\hat{R}[P^{te}, \theta, l(x, y; \theta)] = \frac{1}{m} \sum_{i=1}^{m} \beta^*(y_i^{tr}) \cdot l(x_i^{new}, y_i^{tr}; \theta) \qquad (18)$$

# Correction for Target Shift (TarS, $P_{X|Y}^{te} = P_{X|Y}^{tr}$)

Assumptions:

$\mathbf{A}_1^{\text{TarS}}$ : $P_{X|Y}^{te} = P_{X|Y}^{tr}$ and $P_Y^{te} = P_Y^{tr}$, i.e., $\gamma^*(x, y) \equiv 1$ (the difference between $P_{XY}^{tr}$ and $P_{XY}^{te}$) is caused by a shift in target distribution $P_Y$.

$\mathbf{A}_2^{\text{TarS}}$ : The support of $P_Y^{te}$ is contained in the support of $P_Y^{tr}$ (i.e., roughly speaking, the training set is richer than the test set).

$\mathbf{A}_3^{\text{TarS}}$ : There exists only one possible distribution of $Y$ that, together with $P_{X|Y}^{tr}$, leads to $P_X^{te}$.

$\mathbf{A}_4^{\text{TarS}}$ : Product kernel $kl$ on $\mathcal{X} \times \mathcal{Y}$ is characteristic. (For characteristic kernels, the kernel mean map $\mu$ from the space of distribution to the RKHS is injective, meaning that all information of the distribution is preserved.)

# Correction for Target Shift (TarS, $P_{X|Y}^{te} = P_{X|Y}^{tr}$)

We can draw a biased sample from the training data; here the selection variable depends only on $Y$.

Denote by $P^{new}(\cdot)$ the distribution on this sample. We can make $P_X^{new}$ identical to $P_X^{te}$ by adjusting $P_Y^{new}$ ($P_{X|Y}^{new} = P_{X|Y}^{tr} = P_{X|Y}^{te}$).

Let $P_Y^{new} = \beta(y) \cdot P_Y^{tr}$. To make $P_X^{new}$ identical to $P_X^{te}$, we can adjust $\beta(y)$ to minimize

$$\mathcal{D}(P_X^{te}, P_X^{new}) = \mathcal{D}(P_X^{te}, \int P_Y^{tr} \cdot \beta(y) \cdot P_{X|Y}^{te} dy) \qquad (19)$$

where $\mathcal{D}$ measures the difference between two distributions.

# Correction for Target Shift (TarS, $P_{X|Y}^{te} = P_{X|Y}^{tr}$)

### A Kernel Mean Matching Approach:

Using kernel embedding of conditional as well as marginal distributions, the proposed approaches avoid distribution estimation, and are applicable for high-dimensional problems.

### Marginal distribution:

The kernel mean embedding of $P_X$ is a point in the RKHS given by

$$\mu[P_X] = \mathbb{E}_{X \sim P_X}[\psi(X)] \qquad (20)$$

and its empirical estimate is

$$\hat{\mu}[P_X] = \frac{1}{m} \sum_{i=1}^{m} \psi(x_i) \qquad (21)$$

### Conditional distribution:

The embedding of $P_{X|Y}$ can be considered as an operator mapping from $\mathcal{G}$ to $\mathcal{F}$, defined as

$$\mathcal{U}[P_{X|Y}] = \mathcal{C}_{XY} \mathcal{C}_{YY}^{-1} \qquad (22)$$

where $\mathcal{C}_{XY}$ and $\mathcal{C}_{YY}$ denote the (uncentered) cross-covariance and co-variance operators, respectively.

# Correction for Target Shift (TarS, $P_{X|Y}^{te} = P_{X|Y}^{tr}$)

The kernel mean embedding of $P_Y^{new}$ is

$$\mu[P_Y^{new}] = \mathbb{E}_{Y \sim P_Y^{new}}[\phi(Y)] = \mathbb{E}_{Y \sim P_Y^{tr}}[\beta(y)\phi(Y)] \tag{23}$$

where $\mathcal{D}$ measures the difference between two distributions.
The embedding of $P_X^{new}$ is then given by $\mu[P_X^{new}] = \mathcal{U}[P_{X|Y}^{tr}]\mu[P_Y^{new}]$.
We can find $\beta(y)$ by minimizing the maximum mean discrepancy:

$$\begin{aligned}
||\mu[P_X^{new}] - \mu[P_X^{te}]|| &= \left|\left|\mathcal{U}[P_{X|Y}^{tr}]\mu[P_Y^{new}] - \mu[P_X^{te}]\right|\right| \\
&= \left|\left|\mathcal{U}[P_{X|Y}^{tr}]\mathbb{E}_{Y \sim P_Y^{tr}}[\beta(y)\phi(Y)] - \mu[P_X^{te}]\right|\right|
\end{aligned} \tag{24}$$

subject to $\beta(y) \geq 0$ and $\mathbb{E}_{P_Y^{tr}}[\beta(y)] = 1$, which guarantees that
$P_Y^{new} = \beta(y)P_Y^{tr}$ is a valid distribution.

# Correction for Target Shift (TarS, $P_{X|Y}^{te} = P_{X|Y}^{tr}$)

Let $\beta$ stand for $\beta(y^{tr})$ and $\beta_i$ for $\beta(y_i^{tr})/$ Denote by $\mathbf{1}_n$ the vector of 1's of length $n$, and by $K^c$ the cross kernel matrix between $y^{te}$ and $y^{tr}$, i.e., $K_{ij}^c = \kappa(x_i^{tr}, x_j^{tr})$. The empirical version of the square of (24) is

$$
\begin{aligned}
&\left\| \hat{\mathcal{U}}_{X|Y} \cdot \frac{1}{m} \sum_{i=1}^{m} \beta_i \phi(y_i^{tr}) - \frac{1}{n} \sum_{i=1}^{n} \psi(x_i^{te}) \right\|^2 \\
=&\frac{1}{m^2} \beta^T \phi^T(y^{tr}) \hat{\mathcal{U}}_{X|Y}^T \hat{\mathcal{U}}_{X|Y} \phi(y^{tr}) \beta \\
&\quad - \frac{2}{mn} \mathbf{1}_n^T \psi^T(x^{te}) \hat{\mathcal{U}}_{X|Y} \phi(y^{tr}) \beta + \texttt{const} \\
=&\frac{1}{m^2} \beta^T \underbrace{\Omega K \Omega^T}_{\triangleq A} \beta - \frac{2}{mn} \underbrace{\mathbf{1}_n^T K^c \Omega^T}_{\triangleq M} \beta + \texttt{const}
\end{aligned}
\tag{25}
$$

where the empirical estimate of $\mathcal{U}_{X|Y}$ is $\hat{\mathcal{U}}_{X|Y} = \Psi(L + \lambda I)^{-1} \Phi^T$, and $\Omega \triangleq L(L + \lambda I)^{-1}$.

# Correction for Target Shift (TarS, $P_{X|Y}^{te} = P_{X|Y}^{tr}$)

Therefore, we have the following constrained quadratic programming (QP) problem:

$$\min_{\beta} \frac{1}{2}\beta^T A\beta - \frac{m}{n}M\beta, \text{ s.t. } \beta_i \in [0, B_\beta] \text{ and } \left|\sum_{i=1}^{m}\beta_i - m\right| \leq m\epsilon \qquad (26)$$

where the empirical estimate of $\mathcal{U}_{X|Y}$ is $\hat{\mathcal{U}}_{X|Y} = \Psi(L + \lambda I)^{-1}\Phi^T$, and $\Omega \triangleq L(L + \lambda I)^{-1}$.

$\beta$ values estimated by solving the above optimization problem usually change dramatically along with $y$. To improve the estimation quality of $\beta$, we reparameterize $\beta$ as $\beta = R\alpha$, and obtain:

$$\min_{\beta} \frac{1}{2}\alpha^T R^T A R\alpha - \frac{m}{n}[MR]\alpha,$$
$$\text{s.t. } [R\alpha]_i \in [0, B_\beta] \text{ and } \left|\sum_{i=1}^{m}\mathbf{1}_m^T R\alpha - m\right| \leq m\epsilon \qquad (27)$$

# Location-Scale Conditional Shift (ConS, $P_Y^{te} = P_Y^{tr}$)

### Assumptions:

$\mathbf{A}_1^{\mathbf{ConS}}$ : There exists $\mathbf{w}(Y^{tr}) = \mathtt{diag}[w_1(Y^{tr}), \cdots, w_d(Y^{tr})]$ and $\mathbf{b}(Y^{tr}) = [b_1(Y^{tr}), \cdots, b_d(Y^{tr})]^T$, where $d$ is the dimensionality of $X$, such that the conditional distribution of $X^{new} \triangleq \mathbf{w}(Y^{tr})X^{tr} + \mathbf{b}(Y^{tr})$ given $Y^{tr}$ is the same as that of $X^{te}$ given $Y^{te}$.

$*$ We term this situation location-scale Cons (LS-ConS). In matrix form. the transformed training points:

$$x^{new} \triangleq x^{tr} \odot \mathbf{W} + \mathbf{B} \tag{28}$$

# Location-Scale Conditional Shift (ConS, $P_Y^{te} = P_Y^{tr}$)

### Assumptions:

$\mathbf{A_2^{ConS}}$ : Set $\{c_{i1}P_{X|Y}^{(\mathbf{w_i},\mathbf{b_i})}(x|y_i) + c_{i2}P_{X|Y}^{(\mathbf{w'_i},\mathbf{b'_i})}(x|y_i); i = 1, \cdots, C\}$ is linearly independent $\forall c_{i1}, c_{i2}(c_{i1}^2 + c_{i2}^2 \neq 0)$, $\mathbf{w_i}, \mathbf{w'_i}(||\mathbf{w_i}||_F^2 + ||\mathbf{w'_i}||_F^2 \neq \mathbf{0})$, and $\mathbf{b_i}, \mathbf{b'_i}$.

$*$ A necessary condition for $\mathbf{A_2^{ConS}}$ is that $P_{X|Y}^{tr}(x|y_i)$, $i = 1, \cdots, C$, are linearly independent after any LS transformation.

# Location-Scale Conditional Shift (ConS, $P_Y^{te} = P_Y^{tr}$)

## A Kernel Mean Matching Approach:

We parameterize **W** and **B** as $\mathbf{W} = R\mathbf{G}$ and $\mathbf{B} = R\mathbf{H}$, where **G** and **G** are the parameters to be estimated.

$$
\begin{aligned}
\mathcal{U}[P_{X|Y}^{new}] &= \mathcal{C}_{X^{new}Y}\mathcal{C}_{YY}^{-1} \\
&= \mathbb{E}_{(X^{new},Y)\sim P_{XY}^{new}}\left[\psi(X^{new}) \otimes \phi^T(Y)\right]\mathbb{E}_{Y\sim P_Y^{tr}}^{-1}\left[\phi(Y) \otimes \phi^T(Y)\right] \quad (29) \\
&= \mathbb{E}_{(X^{tr},Y)\sim P_{XY}^{tr}}\left[\psi(X^{new}) \otimes \phi^T(Y)\right]\mathbb{E}_{Y\sim P_Y^{tr}}^{-1}\left[\phi(Y) \otimes \phi^T(Y)\right]
\end{aligned}
$$

The empirical estimate of $\mathcal{U}[P_{X|Y}^{new}]$ is consequently

$$
\begin{aligned}
\hat{\mathcal{U}}[P_{X|Y}^{new}] &= \frac{1}{m}\psi(x^{new}) \cdot \phi^T(y^{tr}) \cdot \left[\frac{1}{m}\phi(y^{tr})\phi^T(y^{tr}) + \tilde{\lambda}I\right]^{-1} \\
&= \tilde{\Psi})(L + \lambda I)^{-1}\Phi^T
\end{aligned}
\quad (30)
$$

where $\tilde{\Psi}) = \psi(x^{new})$.

# Location-Scale Conditional Shift (ConS, $P_Y^{te} = P_Y^{tr}$)

We aim to minimize $||\mu[P_X^{new}] - \mu[P_X^{te}]||^2$, whose empirical version is :

$$
\begin{aligned}
\mathcal{J}^{ConS} &\triangleq ||\hat{\mu}[P_X^{new}] - \hat{\mu}[P_X^{te}]||^2 = \left|\left| \hat{\mathcal{U}}[P_{X|Y}^{new}]\hat{\mu}[P_Y^{tr}] - \hat{\mu}[P_X^{te}] \right|\right|^2 \\
&= \frac{1}{m^2}\mathbf{1}_m^T\phi^T(y^{tr})\hat{\mathcal{U}}^T[P_{X|Y}^{new}]\hat{\mathcal{U}}[P_{X|Y}^{new}]\phi(y^{tr})\mathbf{1}_m \\
&\quad - \frac{2}{mn}\mathbf{1}_n^T\psi^T(x^{te})\hat{\mathcal{U}}[P_{X|Y}^{new}]\phi(y^{tr})\mathbf{1}_m \\
&= \frac{1}{m^2}\mathbf{1}_m^T\Omega\tilde{K}\Omega^T\mathbf{1}_m - \frac{2}{mn}\mathbf{1}_n^T\tilde{K}^c\Omega^T\mathbf{1}_m.
\end{aligned}
\tag{31}
$$

We also regularize (31) to prefer the change in $P_{X|Y}$ to be as little as possible, i.e., to make the entries of **W** close to one and those of **B** close to zero. The regularization term is:

$$\mathcal{J}^{reg} = \frac{\lambda_{LS}}{m} \cdot \left|\left|\mathbf{W} - \mathbf{1}_m \mathbf{1}_d^T\right|\right|_F^2 + \frac{\lambda_{LS}}{m} \cdot \left|\left|\mathbf{B}\right|\right|_F^2 \qquad (32)$$

# LS Generalized Target Shift (GeTarS, $P_Y$, $P_{X|Y}$ change)

Assumptions: Assume $\mathbf{A_1^{ConS}}$ holds, i.e., we condider LS-GeTarS.

We find the empirical version of $||\mu[P_X^{new}] - \mu[P_X^{te}]||^2$

$$
\begin{aligned}
\mathcal{J} &= ||\hat{\mu}[P_X^{new}] - \hat{\mu}[P_X^{te}]||^2 = \left|\left| \hat{\mathcal{U}}[P_{X|Y}^{new}]\hat{\mu}[P_Y^{tr}] - \hat{\mu}[P_X^{te}] \right|\right|^2 \\
&= \left|\left| \frac{1}{m}\hat{\mathcal{U}}[P_{X|Y}^{new}]\phi(y^{tr})\beta - \frac{1}{n}\psi(x^{te})\mathbf{1}_n \right|\right|^2 \\
&= \frac{1}{m^2}\beta^T \Omega \tilde{K} \Omega^T \beta - \frac{2}{mn}\mathbf{1}_n^T \tilde{K}^c \Omega^T \beta.
\end{aligned}
\tag{33}
$$

We would also like the difference between $P_{X|Y}^{te}$ and $P_{X|Y}^{tr}$ to be as little as possible, and minimize:

$$
\mathcal{J}^{GeTarS} = \mathcal{J} + \lambda_{LS}\mathcal{J}^{reg}
$$

# Outline

## Background

The invariant components (IC)-type methods assume that there exist a transformation $\mathcal{T}$ such that $P^S(\mathcal{T}(X)) = P^T(\mathcal{T}(X))$.

Drawback #1: In unsupervised domain adaptation, $\mathcal{T}$ can no be learned by minimizing the distance between $P^S(Y|\mathcal{T}(X))$ and $P^T(Y|\mathcal{T}(X))$. So, previous methods simply assume $P^S(Y|\mathcal{T}(X)) \approx P^T(Y|\mathcal{T}(X))$.

Drawback #2: GeTars method assumes that all the features can be transferred to the target domain by location-scale (LS) transformation. However, many of the features can be highly noisy or cannot be well matched after LS transformation.

Rethinking #1: Under what conditions would $P^S(\mathcal{T}(X)) \approx P^T(\mathcal{T}(X))$ imply $P^S(Y|\mathcal{T}(X)) \approx P^T(Y|\mathcal{T}(X))$?

Rethinking #2: The components that are transferable between domains are not necessarily invariant.

---

[1] Gong et al. Domain Adaptation with Conditional Transferable Components, ICML 2016.

## Causal Mechanism

Approach: Capture the underlying causal mechanism, and use causal models to characterize how the distribution changes between domains.

Causal system: Let $P(C)$ characterizes the process which generates the cause, and $P(E|C)$ describes the mechanism transforming cause $C$ to effect $C$. In the causal system $C \longrightarrow E$, $P(E|C)$ is *independent* of the cause generating process $P(C)$.

Remark: In a causal system $X \longrightarrow Y$, if $P(Y|X)$ changes across domains, one can hardly correct $P(Y|X)$ because $P(X)$ contains no information about $P(Y|X)$.

## Basic Idea

Find transferable components whose conditional distribution is invariant after proper LS transformations, i.e., $P^S(\mathcal{T}(X)|Y) \approx P^T(\mathcal{T}(X)|Y)$.

The causal direction is $Y \longrightarrow X$. Here, $P(Y)$ and $P(X|Y)$ change independently to each other, whereas $P(X)$ and $P(Y|X)$ usually change dependently (thus it is possible to correct $P(Y|X)$).

Conditional invariant components (CIC) $X^{ci}$ are those components satisfying the condition that $P(X^{ci}|Y)$ stays invariant across different domains.

Conditional transferable components (CTC) method: the conditional distribution of the extracted conditional transferable components $X^{ct}$ given $Y$, $P(X^{ct}|Y)$, differs only in the location and scale across all domains.

## Conditional Invariant Components

We first assume that there exist $d$-dimensional conditional invariant components that can be represented as a linear transformation of the $D$-dimensional raw features:

$$X^{ci} = W^T X \tag{34}$$

where $W \in \mathbb{R}^{\mathbb{D} \times}$ is a orthonormal matrix. If we have two domains on which both $X$ and $Y$ are known, we can directly enforce the condition:

$$P^{\mathcal{T}}(X^{ci}|Y) = P^{\mathcal{S}}(X^{ci}|Y) \tag{35}$$

However, in unsupervised domain adaptation, only the empirical marginal distribution of $X$ is available on the test domain. We do not have access to the $Y$ values on the target domain, and cannot match the conditional distributions directly.

## Conditional Invariant Components

### Assumptions:

$\mathbf{A^{CIC}}$ : The elements in the set $\{\kappa_{c1} P^{\mathcal{S}}(W^T X | Y = v_c) + \kappa_{c2} P^{\mathcal{T}}(W^T X | Y = v_c); c = 1, \cdots, C\}$ are linearly independent $\forall \kappa_{c1}, \kappa_{c2} (\kappa_{c1}^2 + \kappa_{c2}^2 \neq 0)$, if ther are not zero.

$\mathbf{A}$ : The linear transformation $W$ is non-trivial.

### Theorem:

If $P^{new}(X^{ci}) = P^{\mathcal{T}}(X^{ci})$, we have $P^{\mathcal{S}}(X^{ci}|Y) = P^{\mathcal{T}}(X^{ci}|Y)$ and $P^{new}(Y) = P^{\mathcal{T}}(Y)$, i.e., $X^{ci}$ are conditional invariant components from the source to the target domain.

## Conditional Invariant Components

A Squared Maximum Mean Discrepancy (MMD) Approach:

$$
\begin{aligned}
\mathcal{J}^{ci} &= \left|\left| \mu_{P^{new}(X^{ci})}[\psi(X^{ci})] - \mu_{P^{\mathcal{T}}(X^{ci})}[\psi(X^{ci})] \right|\right|^2 \\
&= \left|\left| \mathbb{E}_{X^{ci} \sim P^{new}(X^{ci})}[\psi(X^{ci})] - \mathbb{E}_{X^{ci} \sim P^{\mathcal{T}}(X^{ci})}[\psi(X^{ci})] \right|\right|^2 \quad (36) \\
&= \left|\left| \mathbb{E}_{(Y,X) \sim P^{\mathcal{S}}}[\beta(Y)\psi(W^T X)] - \mathbb{E}_{X \sim P^{\mathcal{T}}}[\psi(W^T X)] \right|\right|^2
\end{aligned}
$$

In practice, we minimize its empirical version:

$$
\begin{aligned}
\hat{\mathcal{J}}^{ci} &= \left|\left| \frac{1}{n^{\mathcal{S}}}\psi(W^T x^{\mathcal{S}})\beta - \frac{1}{n^{\mathcal{T}}}\psi(W^T x^{\mathcal{T}})\mathbf{1} \right|\right|^2 \\
&= \frac{1}{n^{\mathcal{S}^2}}\beta^T K_W^{\mathcal{S}}\beta - \frac{2}{n^{\mathcal{S}}n^{\mathcal{T}}}\mathbf{1}^T K_W^{\mathcal{T},\mathcal{S}}\beta + \frac{1}{n^{\mathcal{T}^2}}\mathbf{1}^T K_W^{\mathcal{T}}\mathbf{1}.
\end{aligned} \quad (37)
$$

# Location-Scale Conditional Transferable Components

One may not find sufficient conditional invariant components. Therefore, to find more useful conditional transferable components, we assume that there exist transferable components that can be approximated by a location-scale transformation across domains.

## Assumptions:

We assume that there exists $W$, $\mathbf{a}(Y^{\mathcal{S}}) = [a_1(Y^{\mathcal{S}}), \cdots, a_d(Y^{\mathcal{S}})]^T$ and $\mathbf{b}(Y^{\mathcal{S}}) = [b_1(Y^{\mathcal{S}}), \cdots, b_d(Y^{\mathcal{S}})]^T$, such that the conditional distribution of $X^{ct} \triangleq \mathbf{a}(Y^{\mathcal{S}}) \circ (W^T X^{\mathcal{S}}) + \mathbf{b}(Y^{\mathcal{S}})$ given $Y^{\mathcal{S}}$ is close to that of $W^T X^{\mathcal{T}}$ given $Y^{\mathcal{T}}$.

$*$ The transformed training data matrix can be written in matrix form:

$$x^{ct} = \mathbf{A} \circ (W^T x^{\mathcal{S}}) + \mathbf{B} \tag{38}$$

where $\circ$ denote the Hadamard product.

## Location-Scale Conditional Transferable Components

With (38), we can generalize $\mathcal{J}^{ci}$ to

$$\mathcal{J}^{ct} = \left|\left|\mathbb{E}_{(Y,X^{ct})\sim P^{\mathcal{S}}}[\beta(Y)X^{ct}] - \mathbb{E}_{X\sim P^{\mathcal{T}}}[\psi(W^T X)]\right|\right|^2 \tag{39}$$

and its empirical version:

$$\begin{aligned}
\hat{\mathcal{J}}^{ct} &= \left|\left|\frac{1}{n^{\mathcal{S}}}\psi(x^{ct})\beta - \frac{1}{n^{\mathcal{T}}}\psi(W^T x^{\mathcal{T}})\mathbf{1}\right|\right|^2 \\
&= \frac{1}{n^{\mathcal{S}^2}}\beta^T \tilde{K}^{\mathcal{S}}\beta - \frac{2}{n^{\mathcal{S}}n^{\mathcal{T}}}\mathbf{1}^T \tilde{K}^{\mathcal{T},\mathcal{S}}\beta + \frac{1}{n^{\mathcal{T}^2}}\mathbf{1}^T K^{\mathcal{T}}\mathbf{1}.
\end{aligned} \tag{40}$$

In practice, we add a regularization term on **A** and **B**:

$$\mathcal{J}^{reg} = \frac{\lambda_S}{n^{\mathcal{S}}}\cdot\left|\left|\mathbf{A} - \mathbf{1}_{d\times n^{\mathcal{S}}}\right|\right|_F^2 + \frac{\lambda_L}{n^{\mathcal{S}}}\cdot\|\mathbf{B}\|_F^2$$

.

## Target Information Preservation

We would like $X^{ct}$ preserve the information about $Y$, i.e.,

$$\Sigma_{YY|X^{ct}} - \Sigma_{YY|X} = 0 \tag{41}$$

where $\Sigma_{YY|X}$ is the conditional covariance operator. According to its definition, we have

$$\Sigma_{YY|X^{ct}} = \Sigma_{YY} - \Sigma_{Y,X^{ct}}\Sigma_{X^{ct},X^{ct}}^{-1}\Sigma_{X^{ct},Y} \tag{42}$$

where $\Sigma_{.,.}$ is the covariance or cross-covariance operator. We can use $\frac{1}{n^S}\phi(y^S)\phi^T(y^S)$, $\frac{1}{n^S}\phi(y^S)\psi^T(x^{ct})$, and $\frac{1}{n^S}\psi(x^{ct})\psi^T(x^{ct})$ as the estimators of $\Sigma_{YY}$, $\Sigma_{Y,X^{ct}}$, and $\Sigma_{X^{ct},X^{ct}}$.

*The End!*