Cross-Domain Object Matching with Model Selection

Makoto Yamada
Tokyo Institute of Technology
yamada@sg.cs.titech.ac.jp

Masashi Sugiyama
Tokyo Institute of Technology
sugi@cs.titech.ac.jp

Abstract

The goal of *cross-domain object matching* (CDOM) is to find correspondence between two sets of objects in different domains in an unsupervised way. Photo album summarization is a typical application of CDOM, where photos are automatically aligned into a designed frame expressed in the Cartesian coordinate system. CDOM is usually formulated as finding a mapping from objects in one domain (photos) to objects in the other domain (frame) so that the pairwise dependency is maximized. A state-of-the-art CDOM method employs a kernel-based dependency measure, but it has a drawback that the kernel parameter needs to be determined manually. In this paper, we propose alternative CDOM methods that can naturally address the model selection problem. Through experiments on image matching, unpaired voice conversion, and photo album summarization tasks, the effectiveness of the proposed methods is demonstrated.

1 Introduction

The objective of *cross-domain object matching* (CDOM) is to match two sets of objects in different domains. For instance, in photo album summarization, photos are automatically assigned into a designed frame expressed in the Cartesian coordinate system. A typical approach of CDOM is to find a mapping from objects in one domain (photos) to objects in the other domain (frame) so that the pairwise dependency is maximized. In this scenario, accurately evaluating the dependence between objects is a key challenge.

Appearing in Proceedings of the 14^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL, USA. Volume 15 of JMLR: W&CP 15.

Kernelized sorting (KS) (Jebara, 2004) tries to find a mapping between two domains that maximizes the *mutual information* (MI) (Cover and Thomas, 2006) under the Gaussian assumption. However, since the Gaussian assumption may not be fulfilled in practice, this method (which we refer to as KS-MI) tends to perform poorly.

To overcome the limitation of KS-MI, Quadrianto *et al.* (2010) proposed using the kernel-based dependence measure called the *Hilbert-Schmidt independence criterion* (HSIC) (Gretton *et al.*, 2005) for KS. Since HSIC is distribution-free, KS with HSIC (which we refer to as KS-HSIC) is more flexible than KS-MI. However, HSIC includes a tuning parameter (more specifically, the Gaussian kernel width), and its choice is crucial to obtain better performance (see also Jagarlamudi *et al.*, 2010). Although using the median distance between sample points as the Gaussian kernel width is a common heuristic in kernel-based dependence measures (see e.g., Fukumizu *et al.*, 2009a), this does not always perform well in practice.

In this paper, we propose two alternative CDOM methods that can naturally address the model selection problem. The first method employs another kernel-based dependence measure based on the *normalized cross-covariance operator* (NOCCO) (Fukumizu *et al.*, 2009b), which we refer to as KS-NOCCO. The NOCCO-based dependence measure was shown to be asymptotically independent of the choice of kernels. Thus, KS-NOCCO is expected to be less sensitive to the kernel parameter choice, which is an advantage over HSIC.

The second method uses *least-squares mutual information* (LSMI) (Suzuki *et al.*, 2009) as the dependence measure, which is a consistent estimator of the *squared-loss mutual information* (SMI) achieving the optimal convergence rate. We call this method *least-squares object matching* (LSOM). An advantage of LSOM is that cross-validation (CV) with respect to the LSMI criterion is possible. Thus, all the tuning parameters such as the Gaussian kernel width and the regularization parameter can be objectively determined by

CV.

Through experiments on image matching, unpaired voice conversion, and photo album summarization tasks, LSOM is shown to be the most promising approach to CDOM.

2 Problem Formulation

In this section, we formulate the problem of *cross-domain object matching* (CDOM).

The goal of CDOM is, given two sets of samples of the same size, $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$, to find a mapping that well "matches" them.

Let π be a permutation function over $\{1,\ldots,n\}$, and let $\boldsymbol{\Pi}$ be the corresponding permutation indicator matrix, i.e.,

$$\boldsymbol{\Pi} \in \{0,1\}^{n\times n},\ \boldsymbol{\Pi}\mathbf{1}_n = \mathbf{1}_n,\ \text{and}\ \boldsymbol{\Pi}^\top \mathbf{1}_n = \mathbf{1}_n,$$

where $\mathbf{1}_n$ is the n-dimensional vector with all ones and $^\top$ denotes the transpose. Let us denote the samples matched by a permutation π by

$$Z(\boldsymbol{\Pi}) := \{(\boldsymbol{x}_i, \boldsymbol{y}_{\pi(i)})\}_{i=1}^n.$$

The optimal permutation, denoted by $\boldsymbol{\Pi}^*$, can be obtained as the maximizer of the dependency between the two sets $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$:

$$\boldsymbol{\Pi}^* := \operatorname*{argmax}_{\boldsymbol{\Pi}} D(Z(\boldsymbol{\Pi})),$$

where D is some dependence measure.

3 Existing Methods

In this section, we review two existing methods for CDOM, and point out their potential weaknesses.

3.1 Kernelized Sorting with Mutual Information

Kernelized sorting with mutual information (KS-MI) (Jebara, 2004) matches objects in different domains so that MI between matched pairs is maximized. Here, we review KS-MI following alternative derivation provided in Quadrianto *et al.* (2010).

MI is one of the popular dependence measures between random variables. For random variables X and Y, MI is defined as follows (Cover and Thomas, 2006):

$$\mathrm{MI}(Z) := \iint p(\boldsymbol{x},\boldsymbol{y}) \log \frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y},$$

where $p(\boldsymbol{x},\boldsymbol{y})$ denotes the joint density of $\boldsymbol{x}$ and $\boldsymbol{y}$, and $p(\boldsymbol{x})$ and $p(\boldsymbol{y})$ are marginal densities of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. MI is zero if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, and thus it may be used as a dependency measure. Let $H(X)$, $H(Y)$, and $H(X,Y)$ be the entropies of X and Y and the joint entropy of X and Y, respectively:

$$H(X) = -\int p(\boldsymbol{x}) \log p(\boldsymbol{x}) \mathrm{d}\boldsymbol{x},$$

$$H(Y) = -\int p(\boldsymbol{y}) \log p(\boldsymbol{y}) \mathrm{d}\boldsymbol{y},$$

$$H(X,Y) = -\iint p(\boldsymbol{x},\boldsymbol{y}) \log p(\boldsymbol{x},\boldsymbol{y}) \mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y}.$$

Then MI between X and Y can be written as

$$\mathrm{MI}(Z) = H(X) + H(Y) - H(X,Y).$$

Since $H(X)$ and $H(Y)$ are independent of permutation $\boldsymbol{\Pi}$, maximizing MI is equivalent to minimizing the joint entropy $H(X,Y)$. If $p(\boldsymbol{x},\boldsymbol{y})$ is Gaussian with covariance matrix $\boldsymbol{\Sigma}$, the joint entropy is expressed as

$$H(X,Y) = \frac{1}{2}\log|\boldsymbol{\Sigma}| + \text{Const.},$$

where $|\boldsymbol{\Sigma}|$ denotes the determinant of matrix $\boldsymbol{\Sigma}$.

Now, let us assume that $\boldsymbol{x}$ and $\boldsymbol{y}$ are jointly normal in some reproducing Kernel Hilbert Spaces (RKHSs) endowed with joint kernel $K(\boldsymbol{x},\boldsymbol{x}')L(\boldsymbol{y},\boldsymbol{y}')$, where $K(\boldsymbol{x},\boldsymbol{x}')$ and $L(\boldsymbol{y},\boldsymbol{y}')$ are reproducing kernels for $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Then KS-MI is formulated as follows:

$$\min_{\boldsymbol{\Pi}} \log|\boldsymbol{\Gamma}(\boldsymbol{K} \circ (\boldsymbol{\Pi}^\top \boldsymbol{L}\boldsymbol{\Pi}))\boldsymbol{\Gamma}|, \tag{1}$$

where $\boldsymbol{K} = \{K(\boldsymbol{x}_i,\boldsymbol{x}_j)\}_{i,j=1}^n$ and $\boldsymbol{L} = \{L(\boldsymbol{y}_i,\boldsymbol{y}_j)\}_{i,j=1}^n$ are kernel matrices, $\circ$ denotes the Hadamard product (a.k.a. the element-wise product), $\boldsymbol{\Gamma} = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ is the centering matrix, and $\boldsymbol{I}_n$ is the n-dimensional identity matrix.

A critical weakness of KS-MI is the Gaussian assumption, which may not be fulfilled in practice.

3.2 Kernelized Sorting with Hilbert-Schmidt Independence Criterion

Kernelized sorting with Hilbert-Schmidt independence criterion (KS-HSIC) matches objects in different domains so that HSIC between matched pairs is maximized.

HSIC is a kernel-based dependence measure given as follows (Gretton *et al.*, 2005):

$$\mathrm{HSIC}(Z) = \mathrm{tr}(\bar{\boldsymbol{K}}\bar{\boldsymbol{L}}),$$

where $\bar{\boldsymbol{K}} = \boldsymbol{\Gamma}\boldsymbol{K}\boldsymbol{\Gamma}$ and $\bar{\boldsymbol{L}} = \boldsymbol{\Gamma}\boldsymbol{L}\boldsymbol{\Gamma}$ are the centered kernel matrices for $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Note that

smaller HSIC scores mean that X and Y are closer to be independent.

KS-HSIC is formulated as follows (Quadrianto *et al.*, 2010):

$$\max_{\mathbf{\Pi}} \mathrm{HSIC}(Z(\mathbf{\Pi})), \tag{2}$$

where

$$\mathrm{HSIC}(Z(\mathbf{\Pi})) = \mathrm{tr}(\bar{\boldsymbol{K}}\mathbf{\Pi}^\top \bar{\boldsymbol{L}}\mathbf{\Pi}). \tag{3}$$

This optimization problem is called the *quadratic assignment problem* (QAP) (Finke *et al.*, 1987), and it is known to be *NP-hard*. There exists several QAP solvers based on, e.g., simulated annealing, tabu search, and genetic algorithms. However, those QAP solvers are not easy to use in practice since they contain various tuning parameters.

Another approach to solving Eq.(2) based on a *linear assignment problem* (LAP) (Kuhn, 1955) was proposed in Quadrianto *et al.* (2010), which is explained below. Let us relax the permutation indicator matrix $\mathbf{\Pi}$ to take real values:

$$\mathbf{\Pi} \in [0,1]^{n\times n},\ \mathbf{\Pi}\mathbf{1}_n = \mathbf{1}_n,\ \text{and}\ \mathbf{\Pi}^\top \mathbf{1}_n = \mathbf{1}_n. \tag{4}$$

Then, Eq.(3) is convex with respect to $\mathbf{\Pi}$ (see Lemma 7 in Quadrianto *et al.*, 2010), and its lower bound can be obtained using some $\widetilde{\mathbf{\Pi}}$ as follows:

$$\begin{aligned}
&\mathrm{tr}(\bar{\boldsymbol{K}}\mathbf{\Pi}^\top \bar{\boldsymbol{L}}\mathbf{\Pi}) \\
&\geq \mathrm{tr}(\bar{\boldsymbol{K}}\widetilde{\mathbf{\Pi}}^\top \bar{\boldsymbol{L}}\widetilde{\mathbf{\Pi}}) + \langle \mathbf{\Pi} - \widetilde{\mathbf{\Pi}}, \frac{\partial \mathrm{HSIC}(Z(\widetilde{\mathbf{\Pi}}))}{\partial \mathbf{\Pi}} \rangle \\
&= 2\mathrm{tr}(\bar{\boldsymbol{K}}\mathbf{\Pi}^\top \bar{\boldsymbol{L}}\widetilde{\mathbf{\Pi}}) - \mathrm{tr}(\bar{\boldsymbol{K}}\widetilde{\mathbf{\Pi}}^\top \bar{\boldsymbol{L}}\widetilde{\mathbf{\Pi}}),
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product between matrices. Based on the above lower bound, Quadrianto *et al.* (2010) proposed to update the permutation matrix as

$$\mathbf{\Pi}^{\mathrm{new}} = (1-\eta)\mathbf{\Pi}^{\mathrm{old}} + \eta \operatorname*{argmax}_{\mathbf{\Pi}} \mathrm{tr}\left(\mathbf{\Pi}^\top \bar{\boldsymbol{L}}\mathbf{\Pi}^{\mathrm{old}}\bar{\boldsymbol{K}}\right), \tag{5}$$

where $0 < \eta \leq 1$ is a step size. The second term is an LAP subproblem, which can be efficiently solved by using the *Hungarian method* (Kuhn, 1955).

In the original KS-HSIC paper (Quadrianto *et al.*, 2010), a C++ implementation of the Hungarian method provided by Cooper[1] was used for solving Eq.(5); then $\mathbf{\Pi}$ is kept updated by Eq.(5) until convergence.

In this iterative optimization procedure, the choice of initial permutation matrices is critical to obtain a good solution. Quadrianto *et al.* (2010) proposed the following initialization scheme. Suppose the kernel matrices $\bar{\boldsymbol{K}}$ and $\bar{\boldsymbol{L}}$ are rank one, i.e., for some $\boldsymbol{f}$ and $\boldsymbol{g}$, $\bar{\boldsymbol{K}}$ and $\bar{\boldsymbol{L}}$ can be expressed as $\bar{\boldsymbol{K}} = \boldsymbol{f}\boldsymbol{f}^\top$ and $\bar{\boldsymbol{L}} = \boldsymbol{g}\boldsymbol{g}^\top$. Then HSIC can be written as

$$\mathrm{HSIC}(Z(\mathbf{\Pi})) = \|\boldsymbol{f}^\top \mathbf{\Pi}\boldsymbol{g}\|^2. \tag{6}$$

The initial permutation matrix is determined so that Eq.(6) is maximized. According to Theorems 368 and 369 in Hardy *et al.* (1952), the maximum of Eq.(6) is attained when the elements of $\boldsymbol{f}$ and $\mathbf{\Pi}\boldsymbol{g}$ are ordered in the same way. That is, if the elements of $\boldsymbol{f}$ are ordered in the ascending manner (i.e., $f_1 \leq f_2 \leq \cdots \leq f_n$), the maximum of Eq.(6) is attained by ordering the elements of $\boldsymbol{g}$ in the same ascending way. However, since the kernel matrices $\bar{\boldsymbol{K}}$ and $\bar{\boldsymbol{L}}$ may not be rank one in practice, the principal eigenvectors of $\bar{\boldsymbol{K}}$ and $\bar{\boldsymbol{L}}$ were used as $\boldsymbol{f}$ and $\boldsymbol{g}$ in the original KS-HSIC paper (Quadrianto *et al.*, 2010). We call this *eigenvalue-based initialization*.

Since HSIC is a distribution-free dependence measure, KS-HSIC is more flexible than KS-MI. However, a critical weakness of HSIC is that its performance is sensitive to the choice of kernels (Jagarlamudi *et al.*, 2010). A practical heuristic is to use the Gaussian kernel with width set to the median distance between samples (see e.g., Fukumizu *et al.*, 2009a), but this does not always work well in practice.

4 Proposed Methods

In this section, we propose two alternative CDOM methods that can naturally address the model selection problem.

4.1 Kernelized Sorting with Normalized Cross-Covariance Operator

The kernel-based dependence measure based on the *normalized cross-covariance operator* (NOCCO) (Fukumizu *et al.*, 2009b) is given as follows (Fukumizu *et al.*, 2009b):

$$\mathrm{D}_{\mathrm{NOCCO}}(Z) = \mathrm{tr}(\widetilde{\boldsymbol{K}}\widetilde{\boldsymbol{L}}),$$

where $\widetilde{\boldsymbol{K}} = \bar{\boldsymbol{K}}(\bar{\boldsymbol{K}} + n\epsilon\boldsymbol{I}_n)^{-1}$, $\widetilde{\boldsymbol{L}} = \bar{\boldsymbol{L}}(\bar{\boldsymbol{L}} + n\epsilon\boldsymbol{I}_n)^{-1}$, and $\epsilon > 0$ is a regularization parameter. $\mathrm{D}_{\mathrm{NOCCO}}$ was shown to be asymptotically independent of the choice of kernels. Thus, KS with $\mathrm{D}_{\mathrm{NOCCO}}$ (KS-NOCCO) is expected to be less sensitive to the kernel parameter choice than KS-HSIC.

[1] http://mit.edu/harold/www/code.html

The permuted version of $\widetilde{\boldsymbol{L}}$ can be written as

$$\begin{aligned}\widetilde{\boldsymbol{L}}(\boldsymbol{\Pi}) &= \boldsymbol{\Pi}^\top \bar{\boldsymbol{L}}\boldsymbol{\Pi}(\boldsymbol{\Pi}^\top \bar{\boldsymbol{L}}\boldsymbol{\Pi} + n\epsilon \boldsymbol{I}_n)^{-1} \\ &= \boldsymbol{\Pi}^\top \bar{\boldsymbol{L}}(\bar{\boldsymbol{L}} + n\epsilon \boldsymbol{I}_n)^{-1}\boldsymbol{\Pi} \\ &= \boldsymbol{\Pi}^\top \widetilde{\boldsymbol{L}}\boldsymbol{\Pi},\end{aligned}$$

where we used the orthogonality of $\boldsymbol{\Pi}$ (i.e., $\boldsymbol{\Pi}^\top\boldsymbol{\Pi} = \boldsymbol{\Pi}\boldsymbol{\Pi}^\top = \boldsymbol{I}_n$). Thus, the dependency measure for $Z(\boldsymbol{\Pi})$ can be written as

$$\mathrm{D}_{\mathrm{NOCCO}}(Z(\boldsymbol{\Pi})) = \mathrm{tr}(\widetilde{\boldsymbol{K}}\boldsymbol{\Pi}^\top\widetilde{\boldsymbol{L}}\boldsymbol{\Pi}).$$

Since this is essentially the same form as HSIC, a local optimal solution may be obtained in the same way as KS-HSIC:

$$\boldsymbol{\Pi}^{\mathrm{new}} = (1-\eta)\boldsymbol{\Pi}^{\mathrm{old}} + \eta \operatorname*{argmax}_{\boldsymbol{\Pi}} \mathrm{tr}\left(\boldsymbol{\Pi}^\top\widetilde{\boldsymbol{L}}\boldsymbol{\Pi}^{\mathrm{old}}\widetilde{\boldsymbol{K}}\right). \tag{7}$$

However, the property that $\mathrm{D}_{\mathrm{NOCCO}}$ is independent of the kernel choice holds only asymptotically. Thus, with finite samples, $\mathrm{D}_{\mathrm{NOCCO}}$ does still depend on the choice of kernels as well as the regularization parameter ϵ which needs to be manually tuned.

4.2 Least-Squares Object Matching

Next, we propose an alternative method called *least-squares object matching* (LSOM), in which we employ *least-squares mutual information* (LSMI) (Suzuki *et al.*, 2009) as a dependency measure. LSMI is a consistent estimator of the *squared-loss mutual information* (SMI) with the optimal convergence rate. SMI is defined and expressed as

$$\begin{aligned}&\mathrm{SMI}(Z) \\ &= \frac{1}{2}\iint\left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} - 1\right)^2 p(\boldsymbol{x})p(\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \\ &= \frac{1}{2}\iint\left(\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\right)p(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} - \frac{1}{2}.\end{aligned} \tag{8}$$

Note that SMI is the *Pearson divergence* (Pearson, 1900) from $p(\boldsymbol{x},\boldsymbol{y})$ to $p(\boldsymbol{x})p(\boldsymbol{y})$, while the ordinary MI is the *Kullback-Leibler divergence* (Kullback and Leibler, 1951) from $p(\boldsymbol{x},\boldsymbol{y})$ to $p(\boldsymbol{x})p(\boldsymbol{y})$. SMI is zero if and only if $\boldsymbol{x}$ and $\boldsymbol{y}$ are independent, as the ordinary MI. Its estimator LSMI is given as follows (Suzuki *et al.*, 2009) (see Appendix for the derivation of LSMI):

$$\mathrm{LSMI}(Z) = \frac{1}{2}\widehat{\boldsymbol{\alpha}}^\top\widehat{\boldsymbol{h}} - \frac{1}{2},$$

where

$$\begin{aligned}\widehat{\boldsymbol{\alpha}} &= (\widehat{\boldsymbol{H}} + \lambda\boldsymbol{I}_n)^{-1}\widehat{\boldsymbol{h}}, \\ \widehat{\boldsymbol{H}} &= \frac{1}{n^2}(\boldsymbol{K}\boldsymbol{K}^\top)\circ(\boldsymbol{L}\boldsymbol{L}^\top), \\ \widehat{\boldsymbol{h}} &= \frac{1}{n}(\boldsymbol{K}\circ\boldsymbol{L})\boldsymbol{1}_n.\end{aligned}$$

Here, λ (≥ 0) is the regularization parameter. Since cross-validation (CV) with respect to SMI is possible for model selection, tuning parameters in LSMI (i.e., the kernel parameters and the regularization parameter) can be objectively optimized. This is a notable advantage over kernel-based approaches such as KS-HSIC and KS-NOCCO, since the choice of kernels heavily affects the sensitivity of the independence measure in the kernel-based independence measures (Fukumizu *et al.*, 2009a).

Below, we use the following equivalent expression of LSMI:

$$\mathrm{LSMI}(Z) = \frac{1}{2n}\mathrm{tr}\left(\boldsymbol{L}\widehat{\boldsymbol{A}}\boldsymbol{K}\right) - \frac{1}{2}, \tag{9}$$

where $\widehat{\boldsymbol{A}}$ is the diagonal matrix with diagonal elements given by $\widehat{\boldsymbol{\alpha}}$. Note that we used Eq.(73) and Eq.(75) in Minka (2000) for obtaining the above expression.

LSMI for the permuted data $Z(\boldsymbol{\Pi})$ is given by

$$\mathrm{LSMI}(Z(\boldsymbol{\Pi})) = \frac{1}{2n}\mathrm{tr}\left(\boldsymbol{\Pi}^\top\boldsymbol{L}\boldsymbol{\Pi}\widehat{\boldsymbol{A}}_{\boldsymbol{\Pi}}\boldsymbol{K}\right) - \frac{1}{2},$$

where $\widehat{\boldsymbol{A}}_{\boldsymbol{\Pi}}$ is the diagonal matrix with diagonal elements given by $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Pi}}$, and $\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Pi}}$ is given by

$$\begin{aligned}\widehat{\boldsymbol{\alpha}}_{\boldsymbol{\Pi}} &= (\widehat{\boldsymbol{H}}_{\boldsymbol{\Pi}} + \lambda\boldsymbol{I}_n)^{-1}\widehat{\boldsymbol{h}}_{\boldsymbol{\Pi}}, \\ \widehat{\boldsymbol{H}}_{\boldsymbol{\Pi}} &= \frac{1}{n^2}(\boldsymbol{K}\boldsymbol{K}^\top)\circ(\boldsymbol{\Pi}^\top\boldsymbol{L}\boldsymbol{L}^\top\boldsymbol{\Pi}), \\ \widehat{\boldsymbol{h}}_{\boldsymbol{\Pi}} &= \frac{1}{n}\left(\boldsymbol{K}\circ(\boldsymbol{\Pi}^\top\boldsymbol{L}\boldsymbol{\Pi})\right)\boldsymbol{1}_n.\end{aligned}$$

Consequently, LSOM is formulated as follows:

$$\max_{\boldsymbol{\Pi}}\ \mathrm{LSMI}(Z(\boldsymbol{\Pi})).$$

Since this optimization problem is in general NP-hard and is not convex, we simply use the same optimization strategy as KS-HSIC, i.e., for the current $\boldsymbol{\Pi}^{\mathrm{old}}$, the solution is updated as

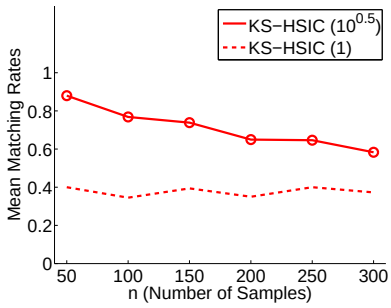
$$\boldsymbol{\Pi}^{\mathrm{new}} = (1-\eta)\boldsymbol{\Pi}^{\mathrm{old}} + \eta \operatorname*{argmax}_{\boldsymbol{\Pi}}\ \mathrm{tr}\left(\boldsymbol{\Pi}^\top\boldsymbol{L}\boldsymbol{\Pi}^{\mathrm{old}}\widehat{\boldsymbol{A}}_{\boldsymbol{\Pi}^{\mathrm{old}}}\boldsymbol{K}\right). \tag{10}$$

5 Experiments

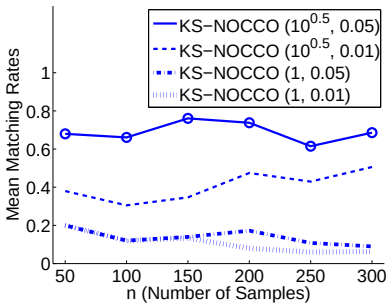
In this section, we experimentally evaluate our proposed algorithms in the image matching, unpaired voice conversion, and photo album summarization tasks.

In all the methods, we use the Gaussian kernels:

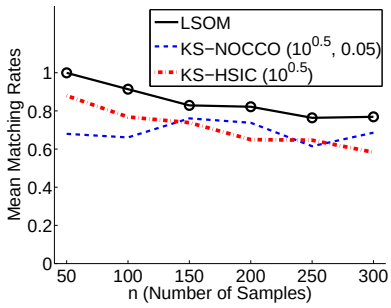
$$\begin{aligned}K(\boldsymbol{x},\boldsymbol{x}') &= \exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\sigma_{\mathrm{x}}^2}\right), \\ L(\boldsymbol{y},\boldsymbol{y}') &= \exp\left(-\frac{\|\boldsymbol{y}-\boldsymbol{y}'\|^2}{2\sigma_{\mathrm{y}}^2}\right),\end{aligned}$$



(a) KS-HSIC with different Gaussian kernel widths.



(b) KS-NOCCO with different Gaussian kernel widths and regularization parameters.



(c) LSOM (tuned by CV), optimally-tuned KS-NOCCO, and optimally-tuned KS-HSIC.

Figure 1: Image matching results. The best method in terms of the mean error and comparable methods according to the t-test at the significance level 1% are specified by '$\circ$'.

and we set the maximum number of iterations for updating permutation matrices to 20 and the step size η to 1. To avoid falling into undesirable local optima, optimization is carried out 10 times with different initial permutation matrices, which are determined by the eigenvalue-based initialization heuristic with Gaussian kernel widths

$$(\sigma_{\mathrm{x}}, \sigma_{\mathrm{y}}) = c \times (m_{\mathrm{x}}, m_{\mathrm{y}}),$$

where $c = 1^{1/2}, 2^{1/2}, \ldots, 10^{1/2}$, and

$$m_{\mathrm{x}} = 2^{-1/2}\mathrm{median}(\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|\}_{i,j=1}^{n}),$$
$$m_{\mathrm{y}} = 2^{-1/2}\mathrm{median}(\{\|\boldsymbol{y}_i - \boldsymbol{y}_j\|\}_{i,j=1}^{n}).$$

In KS-HSIC and KS-NOCCO, we use the Gaussian kernel with the following widths:

$$(\sigma_{\mathrm{x}}, \sigma_{\mathrm{y}}) = c' \times (m_{\mathrm{x}}, m_{\mathrm{y}}),$$

where $c' = 1^{1/2}, 10^{1/2}$. In KS-NOCCO, we use the following regularization parameters:

$$\epsilon = 0.01, 0.05.$$

In LSOM, we choose the model parameters of LSMI, σ_{x}, σ_{y}, and λ by 2-fold CV from

$$(\sigma_{\mathrm{x}}, \sigma_{\mathrm{y}}) = c \times (m_{\mathrm{x}}, m_{\mathrm{y}}),$$
$$\lambda = 10^{-1}, 10^{-2}, 10^{-3}.$$

5.1 Image Matching

Let us consider a toy image matching problem. In this experiment, we use images with RGB format used in Quadrianto *et al.* (2010), which were originally extracted from *Flickr*[2]. We first convert the images from RGB to Lab space and resize them to 40×40 pixels. Next, we convert an image into a 4800-dimensional vector ($4800 = 40 \times 40 \times 3$). Then, we vertically divide images of size 40×40 pixels in the middle, and make two sets of half-images $\{\boldsymbol{x}_i\}_{i=1}^{n}$ and $\{\boldsymbol{y}_i\}_{i=1}^{n}$. Given that $\{\boldsymbol{y}_i\}_{i=1}^{n}$ is randomly permuted, the goal is to recover the correct correspondence.

Figure 2: Image matching result by LSOM. In this case, 234 out of 320 images (73.1%) are matched correctly.

Figure 1 summarizes the average correct matching rate over 100 runs as functions of the number of images, showing that the proposed LSOM method tends to outperform the best tuned KS-NOCCO and KS-NOCCO methods. Note that the tuning parameters of LSOM (σ_{x}, σ_{y}, and λ) are automatically tuned by CV. Figure 2 depicts an example of image matching results obtained by LSOM, showing that most of the images are correctly matched.

[2] http://www.flickr.com

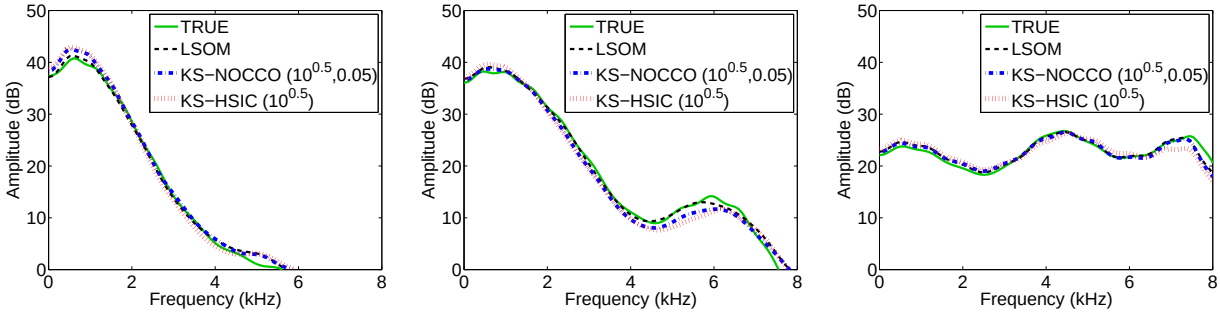


Figure 3: True spectral envelopes and their estimates.

5.2 Unpaired Voice Conversion

Next, we consider an unpaired voice conversion task, which is aimed at matching the voice of a source speaker with that of a target speaker.

In this experiment, we use 200 short utterance samples recorded from two male speakers in French, with sampling rate 44.1kHz. We first convert the utterance samples to 50-dimensional *line spectral frequencies* (LSF) vector (Kain and Macon, 1988). We denote the source and target LSF vectors by $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Then the voice conversion task can be regarded as a multi-dimensional regression problem of learning a function from $\boldsymbol{x}$ to $\boldsymbol{y}$. However, different from a standard regression setup, paired training samples are not available; instead, only unpaired samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$ are given.

By CDOM, we first match $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$, and then we train a multi-dimensional kernel regression model (Schölkopf and Smola, 2002) using the matched samples $\{(\boldsymbol{x}_{\pi(i)}, \boldsymbol{y}_i)\}_{i=1}^n$ as

$$\min_{\boldsymbol{W}} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \boldsymbol{W}^\top \boldsymbol{k}(\boldsymbol{x}_{\pi(i)})\|^2 + \frac{\delta}{2}\mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{W}),$$

where

$$\boldsymbol{k}(\boldsymbol{x}) = (K(\boldsymbol{x}, \boldsymbol{x}_{\pi(1)}), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_{\pi(n)}))^\top,$$
$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|^2}{2\tau^2}\right).$$

Here, τ is a Gaussian kernel width and δ is a regularization parameter; they are chosen by 2-fold CV.

We repeat the experiments 100 times by randomly shuffling training and test samples, and evaluate the voice convergence performance by *log-spectral distance* for 8000 test samples[1] (Quackenbush *et al.*, 1988). Figure 3 shows the true spectral envelope and their estimates, and Figure 4 shows the average performance over 100 runs as the number of training samples. These results show that the proposed LSOM tends to outperform KS-NOCCO and KS-HSIC.

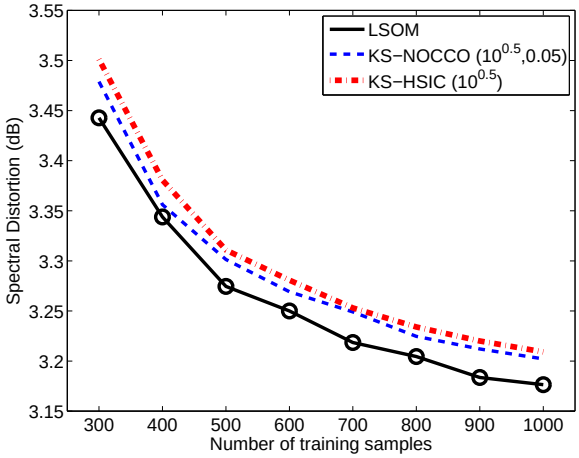


Figure 4: Unpaired voice conversion results. The best method in terms of the mean spectral distortion and comparable methods according to the t-test at the significance level 1% are specified by '○'.

5.3 Photo Album Summarization

Finally, we apply the proposed LSOM method to a photo album summarization problem, where photos are automatically aligned into a designed frame expressed in the Cartesian coordinate system.

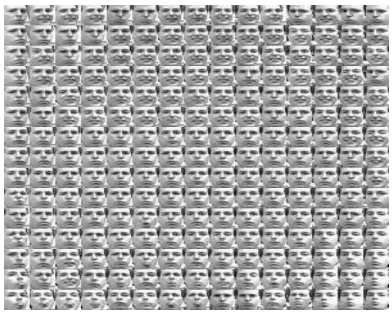
First, we use 320 images in the RGB format obtained from *Flickr*[2]. We consider a rectangular frame of 16×20 ($= 320$), and arrange the images in this rectangular frame. Figure 5(a) depicts the photo album summarization result, showing that images are aligned in the way that images with similar colors are aligned closely.

Similarly, we use the *Frey face dataset* (Roweis and

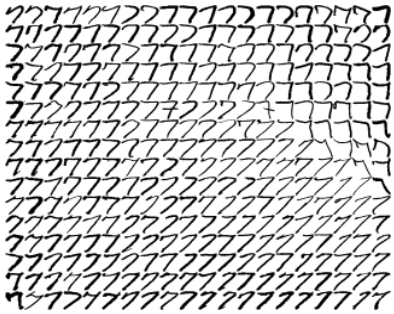
[1] The smaller the spectral distortion is, the better the quality of voice conversion is.

[2] http://www.flickr.com

(a) Layout of 320 images into a 2D grid of size 16 by 20 using LSOM.



(b) Layout of 225 facial images into a 2D grid of size 15 by 15 using LSOM.

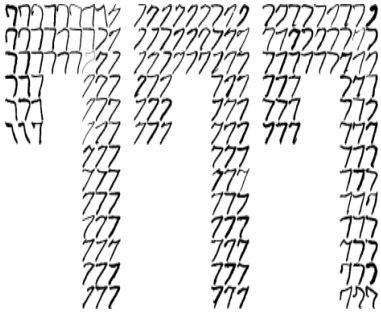


(c) Layout of 320 digit '7' into a 2D grid of size 16 by 20 using LSOM.

Figure 5: Images are automatically aligned into rectangular grid frames expressed in the Cartesian coordinate system.

(a) Layout of 120 images into a Japanese character 'mountain' by LSOM.

(b) Layout of 153 facial images into 'smiley' by LSOM.



(c) Layout of 199 digit '7' into '777' by LSOM.

Figure 6: Images are automatically aligned into complex grid frames expressed in the Cartesian coordinate system.

Saul, 2000), which consists of 225 gray-scale face images with 28×20 ($= 560$) pixels. We similarly convert a image into a 560-dimensional vector, and we set the grid size to 15×15 ($= 225$). The results depicted in Figure 5(b) show that similar face images (in terms of the angle and facial expressions) are assigned in nearby cells in the grid.

Next, we apply LSOM to the USPS hand-written digit dataset (Hastie *et al.*, 2001). In this experiment, we use 320 gray-scale images of digit '7' with 16×16 ($= 256$) pixels. We convert an image into a 256-dimensional vector, and we set the grid size to 16×20 ($= 320$). The result depicted in Figure 5(c) shows that digits with similar profiles are aligned closely.

Finally, we align the Flickr, Frey face, and USPS images into more complex frames—a Japanese character 'mountain', a smiley-face shape, and a '777' digit shape. The results depicted in Figure 6 show that images with similar profiles are located in nearby grid-coordinate cells.

6 Conclusion

In this paper, we proposed two methods of cross-domain object matching (CDOM). The first method uses the dependence measure based on the normalized cross-covariance operator (NOCCO), which is advantageous over HSIC in that NOCCO is asymptotically independent of the choice of kernels. However, with finite samples, it still depends on kernels which need to be manually tuned. To cope with this problem, we proposed a more practical CDOM approach called *least-squares object matching* (LSOM). LSOM adopts *squared-loss mutual information* as a dependence measure, and it is estimated by the method of *least-squares mutual information* (LSMI). A notable advantage of the LSOM method is that it is equipped with a natural cross-validation procedure that allows us to objectively optimize tuning parameters such as the Gaussian kernel width and the regularization parameter in a data-dependent fashion. We applied the proposed methods to the image matching, unpaired voice conversion, and photo album summarization tasks, and experimentally showed that LSOM is the most promising.

Acknowledgments

We thank Dr. Fernando Villavicencio and Dr. Akisato Kimura for their valuable comments. MY acknowledges the JST PRESTO program, and MS acknowledges SCAT, AOARD, and the JST PRESTO program for financial support.

Appendix

SMI cannot be directly computed since it contains unknown densities $p(\boldsymbol{x}, \boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$. Here, we briefly review an SMI estimator called *least-squares mutual information* (LSMI) (Suzuki *et al.*, 2009).

Suppose that we are given n independent and identically distributed (i.i.d.) paired samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ drawn from a joint distribution with density $p(\boldsymbol{x}, \boldsymbol{y})$. A key idea of LSMI is to directly estimate the *density ratio*:

$$w(\boldsymbol{x}, \boldsymbol{y}) = \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})},$$

without going through density estimation of $p(\boldsymbol{x}, \boldsymbol{y})$, $p(\boldsymbol{x})$, and $p(\boldsymbol{y})$.

In LSMI, the density ratio function $w(\boldsymbol{x}, \boldsymbol{y})$ is directly modeled by the following linear model:

$$w_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\ell=1}^{b} \alpha_\ell \varphi_\ell(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\alpha}^\top \boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}), \qquad (11)$$

where b is the number of basis functions, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_b)^\top$ are parameters, and $\boldsymbol{\varphi}(\boldsymbol{x}, \boldsymbol{y}) = (\varphi_1(\boldsymbol{x}, \boldsymbol{y}), \ldots, \varphi_b(\boldsymbol{x}, \boldsymbol{y}))^\top$ are basis functions. Note that, we set $b = n$ in this paper.

The parameter $\boldsymbol{\alpha}$ in the model $w_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y})$ is learned so that the squared error between $w(\boldsymbol{x}, \boldsymbol{y})$ and $w_{\boldsymbol{\alpha}}(\boldsymbol{x}, \boldsymbol{y})$ — this is formulated as

$$\widehat{\boldsymbol{\alpha}} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \Big[\frac{1}{2} \boldsymbol{\alpha}^\top \widehat{\boldsymbol{H}} \boldsymbol{\alpha} - \widehat{\boldsymbol{h}}^\top \boldsymbol{\alpha} + \lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha} \Big],$$

where a regularization term $\lambda \boldsymbol{\alpha}^\top \boldsymbol{\alpha}$ is included for avoiding overfitting, and

$$\widehat{\boldsymbol{H}} = \frac{1}{n^2} \sum_{i,j=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i, \boldsymbol{y}_j) \boldsymbol{\varphi}(\boldsymbol{x}_i, \boldsymbol{y}_j)^\top,$$

$$\widehat{\boldsymbol{h}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\varphi}(\boldsymbol{x}_i, \boldsymbol{y}_i).$$

Here, we use the *product kernel* of the following form as basis functions:

$$\varphi_\ell(\boldsymbol{x}, \boldsymbol{y}) = K(\boldsymbol{x}, \boldsymbol{x}_\ell) L(\boldsymbol{y}, \boldsymbol{y}_\ell),$$

where $K(\boldsymbol{x}, \boldsymbol{x}')$ and $L(\boldsymbol{y}, \boldsymbol{y}')$ are reproducing kernels for $\boldsymbol{x}$ and $\boldsymbol{y}$.

Then $\widehat{\boldsymbol{H}}$ and $\widehat{\boldsymbol{h}}$ can be rewritten as (Petersen and Pedersen, 2008)

$$\widehat{\boldsymbol{H}} = \frac{1}{n^2} (\boldsymbol{K}\boldsymbol{K}^\top) \circ (\boldsymbol{L}\boldsymbol{L}^\top),$$

$$\widehat{\boldsymbol{h}} = \frac{1}{n} (\boldsymbol{K} \circ \boldsymbol{L}) \, \boldsymbol{1}_n.$$

Differentiating the above objective function with respect to $\boldsymbol{\alpha}$ and equating it to zero, we can obtain an analytic-form solution:

$$\widehat{\boldsymbol{\alpha}} = (\widehat{\boldsymbol{H}} + \lambda \boldsymbol{I}_b)^{-1} \widehat{\boldsymbol{h}}.$$

Given a density ratio estimator $\widehat{w} = w_{\widehat{\boldsymbol{\alpha}}}$, SMI can be simply approximated as

$$\mathrm{LSMI}(Z) = \frac{1}{2} \widehat{\boldsymbol{\alpha}}^\top \widehat{\boldsymbol{h}} - \frac{1}{2}.$$

In order to determine the kernel parameter and the regularization parameter λ, cross-validation (CV) is available for the LSMI estimator: First, the samples $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$ are divided into K disjoint subsets $\{\mathcal{S}_k\}_{k=1}^K$, $\mathcal{S}_k = \{(\boldsymbol{x}_{k,i}, \boldsymbol{y}_{k,i})\}_{i=1}^{n_k}$ of (approximately) the same size, where n_k is the number of samples in the subset $\mathcal{S}_k$. Then, an estimator $\widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k}$ is obtained using $\{\mathcal{S}_j\}_{j \neq k}$, and the approximation error for the hold-out samples $\mathcal{S}_k$ is computed as

$$J_{\mathcal{S}_k}^{(K\text{-CV})} = \frac{1}{2} \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k}^\top \widehat{\boldsymbol{H}}_{\mathcal{S}_k} \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k} - \widehat{\boldsymbol{h}}_{\mathcal{S}_k}^\top \widehat{\boldsymbol{\alpha}}_{\mathcal{S}_k},$$

where, for $[\boldsymbol{K}_{\mathcal{S}_k}]_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_{k,j})$, $[\boldsymbol{L}_{\mathcal{S}_k}]_{ij} = L(\boldsymbol{y}_i, \boldsymbol{y}_{k,j})$ $i = 1, \ldots, n, j = 1, \ldots, |\mathcal{S}_k|$,

$$\widehat{\boldsymbol{H}}_{\mathcal{S}_k} = \frac{1}{n_k^2} (\boldsymbol{K}_{\mathcal{S}_k} \boldsymbol{K}_{\mathcal{S}_k}^\top) \circ (\boldsymbol{L}_{\mathcal{S}_k} \boldsymbol{L}_{\mathcal{S}_k}^\top),$$

$$\widehat{\boldsymbol{h}}_{\mathcal{S}_k} = \frac{1}{n_k} (\boldsymbol{K}_{\mathcal{S}_k} \circ \boldsymbol{L}_{\mathcal{S}_k}) \, \boldsymbol{1}_{n_k}.$$

This procedure is repeated for $k = 1, \ldots, K$, and its average $J^{(K\text{-CV})}$ is outputted as

$$J^{(K\text{-CV})} = \frac{1}{K} \sum_{k=1}^{K} J_{\mathcal{S}_k}^{(K\text{-CV})}.$$

We compute $J^{(K\text{-CV})}$ for all model candidates, and choose the model that minimizes $J^{(K\text{-CV})}$.

References

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. John Wiley & Sons, Inc., Hoboken, NJ, USA, 2nd edition.

Finke, G., Burkard, R. E., and Rendl, F. (1987). Quadratic assignment problems. *Annals of Discrete Mathematics*, **31**, 61–82.

Fukumizu, K., Bach, F. R., and Jordan, M. (2009a). Kernel dimension reduction in regression. *The Annals of Statistics*, **37**(4), 1871–1905.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. (2009b). Kernel measures of conditional dependence. In D. Koller, D. Schuurmans, Y. Bengio, and L. Botton, editors, *Advances in Neural Information Processing Systems 21 (NIPS2008)*, pages 489–496, Cambridge, MA. MIT Press.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *16th International Conference on Algorithmic Learning Theory (ALT 2005)*, pages 63–78.

Hardy, G. H., Littlewood, J. E., and Pólya, G. (1952). *Inequalities*. Cambridge University Press, Cambridge.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Jagarlamudi, J., Juarez, S., and Daumé III, H. (2010). Kernelized sorting for natural language processing. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI2010)*, pages 1020–1025, Atlanta, Georgia, U.S.A.

Jebara, T. (2004). Kernelized sorting, permutation, and alignment for minimum volume PCA. In *Conference on Computational Learning theory (COLT)*, pages 609–623.

Kain, A. and Macon, M. W. (1988). Spectral voice conversion for text-to-speech synthesis. In *Proceedings of 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP1998)*, pages 285–288, Washington, DC, U.S.A.

Kuhn, H. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, **2**(1-2), 83–97.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, **22**, 79–86.

Minka, T. P. (2000). Old and new matrix algebra useful for statistics. Technical report, MIT Media Lab.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine*, **50**, 157–175.

Petersen, K. B. and Pedersen, M. S. (2008). The matrix cookbook. Version 20081110.

Quackenbush, S. R., Barnwell, T. P., and Clements, M. A. (1988). *Objective Measures of Speech Quality*. Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Quadrianto, N., Smola, A., Song, L., and Tuytelaars, T. (2010). Kernelized sorting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**, 1809–1821.

Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, **290**, 2323–2326.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. MIT Press, Cambridge, MA.

Suzuki, T., Sugiyama, M., Kanamori, T., and Sese, J. (2009). Mutual information estimation reveals global associations between stimuli and biological processes. *BMC Bioinformatics*, **10**(S52).