Noname manuscript No.
(will be inserted by the editor)

A novel kernel correlation model with the correspondence estimation

Pengwen Chen

Received: date / Accepted: date

Abstract We present a novel multiple linked iterative closest point method to estimate correspondences and the rigid/non-rigid transformations between point-sets/shapes. The estimation task is carried out through maximizing a symmetric similarity function which is the product of square roots of correspondences and a kernel correlation. The local mean square error analysis and robustness analysis are provided to show its superior performance to the kernel correlation method.

Keywords Hellinger distance · Kernel correlation · Monge Kantorovich mass transport problem · point set registration

1 Introduction

Point set representations of image data are commonly used in computer vision and point-sets matching problem is a problem of pivotal importance in these application domains, including medical image alignment[23], segmentation[17], tracking[27], shape recognition[26], and stereo matching[39]. The point pattern matching problem has attracted persistent interest in computer vision[33] and shape statistics[29][8] communities for several decades. The problem might be separated into two main components, the spatial transformation and the point-to-point correspondence. The spatial transformation includes linear transforms(translation, rotation, scaling and affine transforms), and non-rigid transformations(e.g., a variety of splines[3][34], and large deformation models[15][10]). The correspondence problem can be tackled by a variety of methods, e.g., point assignment[25] and spectral methods [20].

Pengwen Chen
Mathematics, University of Connecticut.
E-mail: pengwen@math.uconn.edu

Let $X := \{\mathbf{x}_i : i = 1, \ldots, n_1\}$, the model set and $Y := \{\mathbf{y}_j : j = 1, \ldots, n_2\}$, the scene set be two point-sets belonging to a finite-dimensional vector space $\mathbb{R}^N$, $N = 2$ or 3. Points in a point set are often features extracted from an image, such as the locations of corners, boundary points or salient regions. The task of point-set registration is to establish a consistent point-to-point correspondence between point sets X, Y or to recover the spatial transformation which yields the best alignment. A point set registration method usually should meet several desirable criteria: (1)it can align point sets accurately with a low computational complexity; (2) the alignment should be robust against the existence of outliers or missing corresponding points.

There are extensive studies on this problem, thus we gives a brief review of literature, focusing on those works directly related to this paper. The iterative closest point(ICP) algorithm is one common approach to feature-based image registration problem, because of its simplicity [2] [38]. One limitation of ICP is its local convergence problem which requires sufficient overlap between the data-sets as an initialization. Its vulnerability in performance also includes the proneness to outliers. To alleviate these difficulties, a variety of robust methods were developed. In these models, multiple weighted matches are introduced to reduce the number of local minima. These models can be regarded as multiple-linked ICP methods. In [7], Chui and Rangarjan proposed a robust point matching(RPM) method, which estimates non-rigid transformation and correspondence simultaneously [7]. A technique called soft-assignment is used to establish correspondence according to some probability. Later, the relation between the robust point matching method and the Expectation Maximization(EM) algorithm for Gaussian mixtures was studied in [5]. In [11], Granger and Pennec proposed the EM-ICP method, in which a maximal likelihood approach is used to estimate the transformation and the matches between point sets. In fact, this probabilistic approach has been used in other works, including Wells[36], Luo and Hancock[21], McNeil and Vijayakumar[22], and Tu et al.[32]. They viewed the point sets registration as a maximum likelihood estimation problem. Their optimization algorithm is based on the Expectation Maximization principle: E-step is the computation of probabilities and M-step is the update of the transformation.

Recently Tsin and Kanade [31] proposed a *kernel correlation*(KC) based point-set registration method to align point-sets without establishing the explicit point correspondence: The Gaussian kernel correlation between any two points $\mathbf{x}_i, \mathbf{y}_j$ is given by

$$\mathrm{KC}(\mathbf{x}_i, \mathbf{y}_j) = \exp(-\|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2),$$

and the kernel correlation between point-sets $X := \{\mathbf{x}_i : i = 1, \ldots, n_1\}, Y := \{\mathbf{y}_j : j = 1, \ldots, n_2\}$ is defined as

$$\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \mathrm{KC}(\mathbf{x}_i, \mathbf{y}_j).$$

The matrix with entries $\{\exp(-\|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2) : i = 1, \ldots, n_1; j = 1, \ldots, n_2\}$ is known as the proximity matrix[25], and σ is the *kernel scale*. The transformation T is estimated through $\min_{\mathrm{T}}\{\mathbf{M}_{KC}(T(X), Y)$

$$:= \mathrm{KC}(Y, Y) - 2\mathrm{KC}(Y, \mathrm{T}(X)) + \mathrm{KC}(\mathrm{T}(X), \mathrm{T}(X))\}. \quad (1)$$

Their experiment shows that the multiple-linked registration is more robust than the ICP method, and also outperforms the EM-ICP method in terms of registration accuracy. The KC framework was further improved by Jian and Vemuri in [14]. They regarded points-set as realizations of some mixture of Gaussians, which changes the registration of point-sets into the alignment of two Gaussian mixtures. In this framework, the transformation is estimated through minimizing the L^2 distance between Gaussian distributions. This framework was later applied in the multiple point-sets registration[35].

One goal of this work is to provide some preliminary statistical analysis on the KC method to examine the advantage/disadvantage of establishing correspondences in estimating the unknown transformation. To this end, we introduce a comparable multiple-linked ICP method, called the *Hellinger distance matching*(HD) method, in which a kernel correlation is weighted by square roots of correspondence functions(see Eq.(2)). By maximizing this weighted kernel correlation, both the unknown transformation and correspondence are estimated. This approach can be regarded as a generalization of KC based methods, hence it inherits the robustness of the KC method. However, this model shares several nice and crucial properties: (i) When one point set differs from another point set by a small perturbation, then the correspondence can be identified correctly(we call it the closest point property); (ii) The L^2 Monger-Kantorovich[17] mass transport problem is an asymptotic limit of the HD method when the kernel scale approaches infinity. From this viewpoint, this method reduces the number of local maxima by inhibiting a large number of matching links attached to each individual point.

To clarifies the effect of estimating correspondences, we provide (local) mean square error and robustness analysis for both the KC method and the HD method, which has not done in the existing literatures(Tsin and Kanade [31] pointed out that the influence of the configuration change of each point has exponential decay as a function of distance. But they did not give precise expressions as we provide in this paper.) Usually a mean square error analysis on a point-set matching method could be very complicated. In this paper, we define a novel quantity, *local mean square error(LMSE)*– a fractional rate at which the mean square error increases as noise with an infinitesimal variance is added. The result indicates that due to the sparsity of correspondence matrices, the estimator in the HD method has smaller local mean square errors than that in the KC method. And to measure the robustness against outliers, we take a common tool–*influence functions* from the robustness analysis[13], which measures the infinitesimal estimate bias caused by placing one outlier. The result shows, through identifying correspondences, given the same kernel scale, the HD method has stronger robustness against outliers, i.e., has smaller average shifts. We show that roughly the influence function in the HD method decays twice as fast as that in the KC method. Afterwards, we provide a few experiments, which validates such predictions.

This paper is organized as follows. In section 2, we describe the main idea of the HD method. In section 3, we introduce the associated numerical algorithms, including computing correspondence matrices and an EM-like minimization technique. In section 4, we provide LMSE analysis and robustness analysis. In section 5, we conduct a few experiments to demonstrate the performance of our model. Several lengthy mathematical derivations are listed in the appendix.

1.1 General Notations

- Lowercase bold letters $\mathbf{x}, \mathbf{y}, \ldots$: vectors.
- Uppercase bold letters $\mathbf{A}, \mathbf{M}^{+}, \ldots$: matrices.
- Roman letters $\mathrm{K}, \mathrm{E}, \ldots$: functions.

  Special notations:

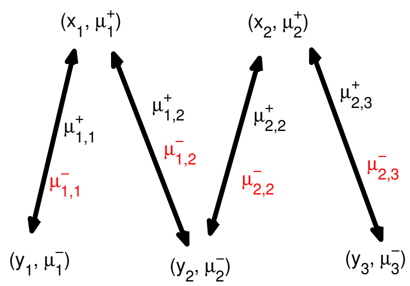

Fig. 1 A matching structure between $X = \{\mathbf{x}_1, \mathbf{x}_2\}$ and $Y = \{\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3\}$.

- X, Y: two point sets.
- N: the dimension of the space which point sets lie in.
- $\mathbf{I}$: the identity matrix.
- $\mathbf{A}^{\dagger}$: the pseudo inverse of $\mathbf{A}$.
- $\mathbf{A}^{\top}$: the transport matrix of $\mathbf{A}$. $\mathfrak{Tr}(\mathbf{A})$: the trace of the square matrix $\mathbf{A}$.

- $\delta_{i,j}$: the Kronecker delta.
- $Normal(\mathbf{c}, b^2)$: a normal distribution with mean $\mathbf{c}$ and variance b^2. $\mathcal{E}[\mathbf{y}]$: the expectation of a random variable $\mathbf{y}$.
  $\mathfrak{Var}(\mathbf{y})$: the variance of a random variable $\mathbf{y}$.

Throughout this paper, $\{a_{i,j} : i, j\}$ refers to a set consisting of elements $\{a_{1,1}, a_{1,2}, \ldots, a_{2,1}, a_{2,2}, \ldots\}$; In other words, when omitting the range of i, j, we mean that i, j go over all possible values.

2 Problem formulation.

The goal of this paper is to present a method to match two point-sets $X := \{\mathbf{x}_i : i = 1, \ldots, n_1\}, Y := \{\mathbf{y}_j : j = 1, \ldots, n_2\}$ given in two Euclidean spaces $\mathbb{R}^N$ ($N = 2$ if in planes). These two point-sets could consist of different number of points due to the existence of outliers. Our task is to estimate a correspondence between the point-sets and a transformation function T between two Euclidean spaces.

Assign a unit weight $\mu_i^+ = 1$ to each point $\mathbf{x}_i$, and a unit weight $\mu_j^- = 1$ to each point $\mathbf{y}_j$(i.e., every point has a unit-weight). A matching structure is described by two nonnegative $n_1 \times n_2$ *matching(correspondence) matrices* $\mathbf{M}^+ := \{\mu^+_{i,j} : i, j\}, \mathbf{M}^- := \{\mu^-_{i,j} : i, j\}$ with the property $1 = \mu_i^+ = \sum_{j=1}^{n_2} \mu^+_{i,j}, 1 = \mu_j^- = \sum_{i=1}^{n_1} \mu^-_{i,j}$ for all i, j. In other words, $\{\mu^+_{i,j} : j\}$ is a partition of $\mu_i^+ = 1$ and $\{\mu^-_{i,j} : i\}$ is a partition of $\mu_j^- = 1$ (see Fig.1).[1]

The matching structure between point-sets is not restricted to one-to-one, i.e., a fuzzy matching is allowed. The matching similarity of the matching matrices $\mathbf{M}^+, \mathbf{M}^-$ between point-sets X, Y is defined as $\mathrm{E}_{HD}(\mathbf{M}^+, \mathbf{M}^-)$

$$= \sum_{i=1,j=1}^{n_1,n_2} \sqrt{\mu^+_{i,j}}\sqrt{\mu^-_{i,j}}\mathrm{K}(\mathbf{x}_i, \mathbf{y}_j), \tag{2}$$

where K is a nonnegative function. Then the similarity of X, Y is given by the maximum

$$\mathrm{E}_{HD}(X, Y) = \max_{\mathbf{M}^+, \mathbf{M}^-} \mathrm{E}_{HD}(\mathbf{M}^+, \mathbf{M}^-),$$

subject to $n_1 + n_2$ marginal constraints:

$$\sum_{j=1}^{n_2} \mu^+_{i,j} = 1, \sum_{i=1}^{n_1} \mu^-_{i,j} = 1,$$

and the matching structure is characterized by its maximizer.[2] Because $\mu^+_{i,j} > 0$ if and only if $\mu^-_{i,j} > 0$ for each pair (i, j)(see A.2.), then a matching structure between X, Y can be described by *matching links*, which are a number of nonzero pairs $\{(\mathbf{x}_i, \mathbf{y}_j) : \mu^+_{i,j} > 0, \mu^-_{i,j} > 0\}$, and each nonzero pair $(\mu^+_{i,j} > 0, \mu^-_{i,j} > 0)$ is called a *matching pair* between $\mathbf{x}_i$ and $\mathbf{y}_j$. Throughout this paper we select $\mathrm{K}(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/\sigma^2)$.

Note that $\mathrm{E}_{HD}(X, X) = \sum_{i=1}^{n_1} \mu_i^+$, $\mathrm{E}_{HD}(Y, Y) = \sum_{j=1}^{n_2} \mu_j^-$, and

$$\mathrm{M}_{ED}(X, Y) := \mathrm{E}_{HD}(X, X) - 2\mathrm{E}_{HD}(X, Y) + \mathrm{E}_{HD}(Y, Y) \tag{4}$$

is a distance function which is always nonnegative and equals to zero if and only if $X = Y$(see B.1.). Thus we estimate the transformation by

$$\max_{\mathrm{T}} \mathrm{E}_{HD}(\mathrm{T}(X), Y) + \text{ some regularity term.} \tag{5}$$

We call the above framework *the Hellinger distance(HD) matching model.*

[1] In this paper, we use $\mu_i^+ = 1 = \mu_j^-$, except in the robustness analysis.

[2] An equivalent formulation of E_{HD} is to maximize

$$\sum_{i=1,j=1}^{n_1,n_2} \gamma^+_{i,j}\gamma^-_{i,j}\mathrm{K}(\mathbf{x}_i, \mathbf{y}_j), \tag{3}$$

with respect to $\{\gamma^+_{i,j}, \gamma^-_{i,j} : i, j\}$ subject to normalized l^2 norm constraints, $\sum_{j=1}^{n_2}(\gamma^+_{i,j})^2 = 1$ and $\sum_{i=1}^{n_1}(\gamma^-_{i,j})^2 = 1$.

Remark 1 Why do we use two correspondence matrices, instead of one? First, the kernel correlation is symmetric, i.e., $\mathrm{KC}(X,Y) = \mathrm{KC}(Y,X)$. Then to have a fair comparison, we require $\mathrm{E}_{HD}(X,Y) = \mathrm{E}_{HD}(Y,X)$. Secondly, correspondence problems could be viewed as point assignment problems with two sets of constraints. This asymmetric setting(in general, $\mathbf{M}^+ \neq \mathbf{M}^-$) has a vantage point to handle these constraints in optimization. When a symmetric setting (one assignment matrix) is used, the correspondence matrix is constrained to be doubly stochastic. Although this requirement can be handled by the Sinkhorn matrix balancing technique[28] as shown in the RPM method, the algorithm might converge rather slowly(it converges linearly with a rate– the sub-dominant singular value of the correspondence matrix[19], quite close to 1 in case that the number of points are large). In contrast, the HD method adopts two stochastic matrices to meet each set of constraints separately, then we avoid this computational burden caused by the doubly stochastic constraint. This type of algorithm is called block coordinate descent method. We shall discuss this issue further in Section 3.

Remark 2 Here we make some comparisons between the HD model(Eq. 4) and the KC model(Eq. 1). In case of rigid motions, only the second terms are computed in both models. However, in case of or non-rigid transformations, $\max_{\mathrm{T}} 2\mathrm{KC}(Y, \mathrm{T}(X)) - \mathrm{KC}(\mathrm{T}(X), \mathrm{T}(X))$ is computed in the KC model, but still only the second term in Eq. (4) is computed in the HD model.

Here is a note about the denomination, HD. Hellinger distance(HD) is a well-known probability distance function: Given two probability density function p, q of two probability measures μ^+, μ^- with respect to a dominating measure λ, the Hellinger distance between them is $(\int(\sqrt{\mathrm{p}}-\sqrt{\mathrm{q}})^2 d\lambda)^{1/2}$, or equivalently $\left(2 - 2\int\sqrt{\mathrm{pq}}d\lambda\right)^{1/2}$. This form of square root function has been used in shape matching problems[37].

2.1 Properties of matching pairs

To understand the advantage of the HD model, we list several properties related to matching pairs.

First, suppose that X, Y have the same cardinality $n = n_1 = n_2$, and $\mathbf{y}_i$ is the closest point of $\mathbf{x}_i$ and $\mathbf{x}_i$ is also the closest point of $\mathbf{y}_i$ for each i, i.e., $\|\mathbf{x}_i - \mathbf{y}_i\| \leq \|\mathbf{x}_i - \mathbf{y}_j\|$ and $\|\mathbf{x}_i - \mathbf{y}_i\| \leq \|\mathbf{x}_j - \mathbf{y}_i\|$ for $j = 1, \ldots, n$. Then each closest point can be identified correctly: the matching links are given by $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \ldots, n\}$, i.e., $\mathbf{M}^+ = \mathbf{I} = \mathbf{M}^-$. In this special case, only n matching pairs exist(see B.2.). We call this property *the closest point property*, which shall play a key role in MSE and robustness analysis in section 4. Many existing methods do not share this closest point property, e.g. [20](see Fig. 2).

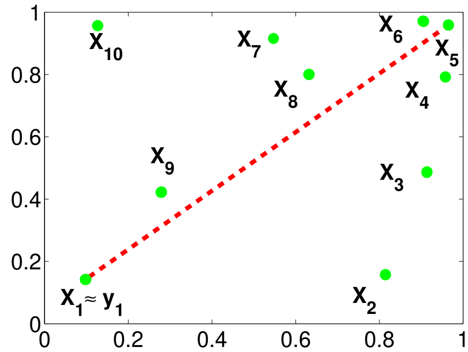

Fig. 2 The spectral technique in [20] does not share the closest point property. Let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_{10}\}$, $Y := \{\mathbf{y}_1, \ldots, \mathbf{y}_{10}\}$ with the relation $\mathbf{y}_i = \mathbf{x}_i$+Normal$(0, .001^2)$ as shown here(since each pair $\mathbf{x}_i, \mathbf{y}_i$ are very adjacent, only $\mathbf{x}_i$ are labeled). The desired matching links are $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \ldots, 10\}$. The matching links are indicated by dashed lines as shown in the figure. All matching links are marked correctly except two wrong assignments $(\mathbf{x}_1, \mathbf{y}_5), (\mathbf{x}_5, \mathbf{y}_1)$. In their work, these incorrect pairs are called accidental assignments.

Secondly, viewing matching pairs as edges in a graph having $n_1 + n_2$ nodes, in general(the proximity matrix has full rank) matching pairs does not form cycles. [3] Thus, there are at most $n_1 + n_2 - 1$ matching pairs(see A.2.), and the optimal correspondence matrices $\mathbf{M}^+, \mathbf{M}^-$ are sparse. On the other hand, there are at least $\max(n_1, n_2)$ matching pairs(from each $\mathbf{x}_i$ and each $\mathbf{y}_j$).
These two properties does not depend on σ. Thirdly, in the asymptotic case $\sigma \to \infty$, then $\mu^+_{i,j}/\mu^-_{i,j} \to 1$ for all i, j, and then the correspondence problem becomes the well-known Monge-Kantorovich mass transport problem[16](see B.3.).[4] The large value σ basically inhibits a large number of matching links attached to one single point through "equalizing the number of matching links attached to each individual point". In contrast, the ICP method tolerates many matching links connecting to a single point, which usually yields a small convergence range(Fig. 3).

[3] For example, a cycle is formed by these 4 matching links: $(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_1), (\mathbf{x}_1, \mathbf{y}_2), (\mathbf{x}_2, \mathbf{y}_2)$ between $X = \{\mathbf{x}_1, \mathbf{x}_2\}$ and $Y = \{\mathbf{y}_1, \mathbf{y}_2\}$.

[4] Roughly speaking, the discrete version of the mass transport problem is to find an optimal permutation p from $\{1, \ldots, n\}$ to itself, so that $\sum_{i=1}^n \mathbf{x}_i \cdot \mathbf{y}_{\mathrm{p}(i)}$ is maximized. More properties in Monge-Kantorovich mass transport problem can be found in [9].

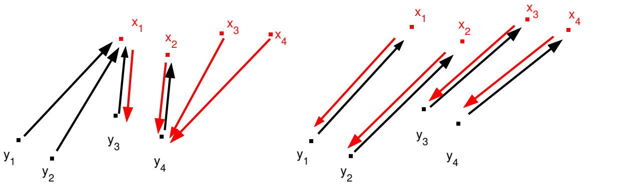

Fig. 3 Illustration of the ICP method stuck at a local optimal solution. Three points $\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3$ share the same closest point $\mathbf{x}_1$ in the ICP method(Left). Large kernel scales in the HD method favor one-to-one correspondences in point-sets in case of equal cardinality (Right).

2.2 Computing correspondence matrices.

Since this cost functional in Eq. (2) is concave both in $\mu^+_{i,j}$ and $\mu^-_{i,j}$, then we can derive the alternating maximizing algorithm to compute the optimal correspondence: The initialization is $\mu^+_{i,j} = 1/n_2$, $\mu^-_{i,j} = 1/n_1$ foe each i, j. The main iteration consists of repeating these two steps,

$$\mu^+_{i,j} \leftarrow \mu^+_i \mu^-_{i,j} \, \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2 / (\sum_{j=1}^{n_2} \mu^-_{i,j} \, \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2), \tag{6}$$

$$\mu^-_{i,j} \leftarrow \mu^-_j \mu^+_{i,j} \, \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2 / (\sum_{i=1}^{n_1} \mu^+_{i,j} \, \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2). \tag{7}$$

The key idea is that given $\mathbf{M}^+$ fixed, $\mathbf{M}^-$ described in Eq. 7 maximizes $\mathrm{E}_{HD}(X, Y)$, and vice versa. The derivation of this algorithm can be found in the appendix(A.1.).

Here we show some differences between the correspondence obtained in our approach and in the soft-assignment [5]. First, obviously two matrices are updated simultaneously here instead of one matrix in the soft-assignment. In the soft-assignment, the limit of the matrix sequence is doubly stochastic (or doubly sub-stochastic if with inequality constraints); in Eq. (6-7), two stochastic matrices are obtained if $\mu^+_i = 1 = \mu^-_j$. Secondly, generally under a fixed kernel scale σ, the correspondence matrix gained in the soft-assignment is not sparse and the matching structure is extracted from the limit as $\sigma \to 0$; on the other hand, the matrices in Eq. (6-7) are highly sparse(usually only $n_1 + n_2 - 1$ nonzero entries) and the nonzero entries directly represents the correspondence between point-sets.

Example 1 Let $\mathbf{x}_1 = 1 = \mathbf{y}_1, \mathbf{x}_2 = 2 = \mathbf{y}_2$ in $\mathbb{R}^1$, $\sigma = 1$, then the optimal correspondence matrix in the soft-assignment is $\begin{pmatrix} 0.73, 0.27 \\ 0.27, 0.73 \end{pmatrix}$; By Eq. (6-7), $\mathbf{M}^+ = \mathbf{M}^- = \begin{pmatrix} 1, 0 \\ 0, 1 \end{pmatrix}$.

Meanwhile, outliers in point-sets can be detected by $\{(i, j) : \mu^+_{i,j} \mu^-_{i,j} < \epsilon\}$.

Example 2 Let $x_1 = 0 = y_1, x_2 = 2, y_2 = -2$, $\sigma = 1$, then by Eq. (6-7), $\mathbf{M}^+ = \begin{pmatrix} 0.98, 0.02 \\ 1, \quad 0 \end{pmatrix}$, $\mathbf{M}^- = \begin{pmatrix} 0.98, 1 \\ 0.02, 0 \end{pmatrix}$. Then the matching link is $\{x_1, y_1\}$, and x_2, y_2 are detected as outliers if $\epsilon = 0.05$. Roughly those terms in E_{HD}, associated with outliers, have less influence in the estimator of the transformation. We shall give more precise descriptions in section 4.

3 Algorithm

In this section, we describe numerical algorithms to solve Eq. (5). First, we describe an EM-like technique to deal with the exponential functions in the above nonlinear optimization problems. With this tool, we can avoid the need of line searches in solving optimization problems, instead we solve a small size($N \times N$) weighted least square problem in each iteration. The main idea is to rewrite the exponential function,

$$\exp(-\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2) = -\min_{r>0}(r \log r - r + r\|\mathbf{x}-\mathbf{y}\|^2/\sigma^2), \tag{8}$$

where the minimum is attained at $r = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/\sigma^2)$. The proof: $r \log r - r$, $\exp(-s)$ are convex conjugates, then $r \log r - r + \exp(-s) \geq rs$, and finally let $s = -\|\mathbf{x} - \mathbf{y}\|^2/\sigma^2$.

We discuss three different cases about T, including affine transformations, rigid motions, and non-rigid motions. In the first two cases, there is no regularization term.

3.1 Weighted least square problems

In the case of affine transformations $\mathrm{T}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{t}$ with $\mathbf{A} \in \mathbb{R}^{N \times N}$, $\mathbf{t} \in \mathbb{R}^{N \times 1}$, plugging Eq. (8) into Eq. (5), the estimation is done through

$$\max_{\mathbf{M}^+, \mathbf{M}^-} \max_{\mathbf{A}, \mathbf{t}} \sum_{i=1, j=1}^{n_1, n_2} \sqrt{\mu^+_{i,j} \mu^-_{i,j}} \exp(-\|\mathbf{A}\mathbf{x}_i + \mathbf{t} - \mathbf{y}_j\|^2/\sigma^2). \tag{9}$$

---

[5] Here we consider "Soft-assignment without inequality constraints"(section 3.2,[7]), which is essentially computed as follows: starting with a matrix $\{\exp(-\|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2) : i, j\}$, perform the row-normalization and column normalization iteratively.

Using Eq. (8), it is obvious to see that Eq. (9) can be solved by repeating three steps:

Alg. 1

Initialization: $\mu^{+}_{i,j} = 1/n_2$, $\mu^{-}_{i,j} = 1/n_1$, $r_{i,j} = 1$.

Main iterations:

1. solve the least square problem: $(\mathbf{A}, \mathbf{t}) = \min_{\mathbf{A},\mathbf{t}} \sum_{i=1,j=1}^{n_1,n_2} s_{i,j} \|(\mathbf{A}\mathbf{x}_i + \mathbf{t}) - \mathbf{y}_j\|^2$, with $s_{i,j} := \sqrt{\mu^{+}_{i,j}\mu^{-}_{i,j}} r_{i,j}$.
2. update $r_{i,j} = \exp(-\|(\mathbf{A}\mathbf{x}_i + \mathbf{t}) - \mathbf{y}_j\|^2/\sigma^2)$.
3. update $\mu^{+}_{i,j}, \mu^{-}_{i,j}$ by one iteration Eq. (6-7), where $\mathbf{x}_i$ is replaced by $\mathbf{A}\mathbf{x}_i + \mathbf{t}$.

The first step can be written explicitly: let

$$\mathbf{t}_y = \frac{\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}\mathbf{y}_j}{\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}}, \mathbf{t}_x = \frac{\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}\mathbf{x}_i}{\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}},$$

and write $\sum_{i=1,j=1}^{n_1,n_2} s_{i,j} \|(\mathbf{A}\mathbf{x}_i + \mathbf{t}) - \mathbf{y}_j\|^2$

$$= \sum_{i=1,j=1}^{n_1,n_2} s_{i,j}(\|\mathbf{A}(\mathbf{x}_i - \mathbf{t}_x) - (\mathbf{y}_j - \mathbf{t}_y)\|^2 + \|\mathbf{A}\mathbf{t}_x - \mathbf{t}_y + \mathbf{t}\|^2$$

$$+2(\mathbf{A}(\mathbf{x}_i - \mathbf{t}_x) - (\mathbf{y}_j - \mathbf{t}_y)) \cdot (\mathbf{A}\mathbf{t}_x - \mathbf{t}_y + \mathbf{t})).$$

$$= \sum_{i=1,j=1}^{n_1,n_2} s_{i,j}(\|\mathbf{A}(\mathbf{x}_i - \mathbf{t}_x) - (\mathbf{y}_j - \mathbf{t}_y)\|^2 + \|\mathbf{A}\mathbf{t}_x - \mathbf{t}_y + \mathbf{t}\|^2).$$

Then the optimal vector $\mathbf{t}$ is given by $\mathbf{t} = \mathbf{t}_y - \mathbf{A}\mathbf{t}_x$, where $\mathbf{A}$ is a solution of a weighted least square equation: let $\mathbf{G}_{yx} = \sum_{i=1,j=1}^{n_1,n_2} s_{i,j}(\mathbf{y}_j - \mathbf{t}_y)(\mathbf{x}_i - \mathbf{t}_x)^\top$, and $\mathbf{G}_{xx} = \sum_{i=1,j=1}^{n_1,n_2} s_{i,j}(\mathbf{x}_i - \mathbf{t}_x)(\mathbf{x}_i - \mathbf{t}_x)^\top$, then $\mathbf{A} = \mathbf{G}_{yx}\mathbf{G}^{+}_{xx}$.

Minor issues: The undesired result $\mathbf{A} = \mathbf{0}$ is obtained at the first iteration, thus instead we set $\mathbf{A} = \mathbf{I}$, and only compute $\mathbf{t}$ at the first iteration. Notice that we do NOT suggest the following alternative algorithm:

– Update correspondence matrices: repeating Eq. (6,7) until it converges.
– Update transformations.

The reason is that this kind of matrix balancing(Eq. (6,7)) usually converges slowly.

How about its convergence property? The proposed algorithm(Alg. 1) is known as block coordinate ascent (descent) method, whose convergence result is ensured by the proposition 2.7.1 in [1], i.e., every limit point of the iterate sequence of $(\mathbf{M}^+, \mathbf{M}^-, \mathbf{A}, \mathbf{b})$ is a stationary point. That is, $(\mathbf{M}^+, \mathbf{M}^-, \mathbf{A}, \mathbf{b})$ in Alg. 1 always converges to a local maximizer.

The complexity of the above is $O(kn_1n_2)$, where k is the number of iterations used to reach the user-defined precision. Empirically, k depends on σ: a smaller k is needed when σ is smaller. To reduce the chance of sticking to local solutions which is not global solutions, we choose σ approximately (or greater than) the average distance among the desired matching pairs.

3.2 Rigid motions

In case of rigid motions, i.e., $\mathrm{T}(\mathbf{x}) = \mathbf{Q}\mathbf{x} + \mathbf{t}$, with $\mathbf{Q}$ being an $N \times N$ orthogonal matrix and $\mathbf{t}$ being an $N \times 1$ vector, we write down explicitly how to solve the associated constrained least square problem(step 1 of Alg. 1.):

$$(\mathbf{Q}, \mathbf{t}) = \min_{\mathbf{Q},\mathbf{t}} \sum_{i=1,j=1}^{n_1,n_2} s_{i,j} \|(\mathbf{Q}\mathbf{x}_i + \mathbf{t}) - \mathbf{y}_j\|^2,$$

with $s_{i,j} := \sqrt{\mu^{+}_{i,j}\mu^{-}_{i,j}} r_{i,j}$. Let

$$\mathbf{t}_y = \frac{\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}\mathbf{y}_j}{\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}}, \mathbf{t}_x = \frac{\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}\mathbf{x}_i}{\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}},$$

and $\mathbf{G} = \sum_{i=1,j=1}^{n_1,n_2} s_{i,j}(\mathbf{x}_i - \mathbf{t}_x)(\mathbf{y}_j - \mathbf{t}_y)^\top$. Compute the singular value decomposition of $\mathbf{G}$, $\mathbf{G} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Then the optimal solution is given by $\mathbf{Q} = \mathbf{V}\mathbf{U}^\top$, and $\mathbf{t} = \mathbf{t}_y - \mathbf{Q}\mathbf{t}_x$. Here is the proof.

Proof Let $\hat{\mathbf{x}}_i = \mathbf{x}_i - \mathbf{t}_x$, $\hat{\mathbf{y}}_j = \mathbf{y}_j - \mathbf{t}_y$, then write $\sum_{i=1,j=1}^{n_1,n_2} s_{i,j} \|(\mathbf{Q}\mathbf{x}_i + \mathbf{t}) - \mathbf{y}_j\|^2$

$$= \sum_{i=1,j=1}^{n_1,n_2} s_{i,j}(\|\mathbf{Q}(\mathbf{x}_i - \mathbf{t}_x) - (\mathbf{y}_j - \mathbf{t}_y)\|^2 + \|\mathbf{Q}\mathbf{t}_x - \mathbf{t}_y + \mathbf{t}\|^2).$$

Then clearly $\mathbf{t} = \mathbf{t}_y - \mathbf{Q}\mathbf{t}_x$ and

$$\mathbf{Q} = arg\max_{\mathbf{Q}}(\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}\|\hat{\mathbf{y}}_j - \mathbf{Q}\hat{\mathbf{x}}_i\|^2)$$

$$= arg\max_{\mathbf{Q}}\{(\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}\hat{\mathbf{y}}_j^\top \mathbf{Q}\hat{\mathbf{x}}_i)$$

$$= \mathrm{Tr}(\mathbf{Q}(\sum_{i=1,j=1}^{n_1,n_2} s_{i,j}\hat{\mathbf{x}}_i\hat{\mathbf{y}}_j^\top))\},$$

Based on the theorem 7.4.9(page 432[12]), $\mathbf{Q}$ is a maximizer if and only if $\mathbf{Q}\mathbf{G}$ is positive semi-definite, which is true($\mathbf{Q}\mathbf{G} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$). □

One thing is worth a mention. At the first iteration, $\mathbf{t}_x, \mathbf{t}_y$ are mass centers of X, Y, but we have $\mathbf{G} = \mathbf{0}$, then $\mathbf{Q}$ is not unique. Hence, we always let $\mathbf{Q} = \mathbf{I}$ at the first iteration, and then $\mathbf{t} = \mathbf{t}_y - \mathbf{t}_x$, the mass centers are matched.

This rigid motion algorithm can be used in the KC method, just letting $s_{i,j} = 1$ and skipping step 3. From our experiment results, this proposed algorithm converges to local maximal in a faster speed than the gradient method suggested in [14].

3.3 Nonrigid motions: combination with splines.

The discussion in this subsection is not new. This kind of combination can be found in many works, e.g. [6] [14]. For the completeness of this paper, we briefly outline the main idea. Spline models have been used widely in estimating a "smooth" transformation function g which assigns $\mathbf{x}_i$ to a point close to $\mathbf{y}_i$ for each i. In the theory of splines, given n control points $\{\mathbf{x}_i : i = 1, \ldots, n\}$, this function g is selected as a minimizer of the functional $\mathrm{E}_{spline}(\mathrm{g}) := \sum_{j=1}^{n} \|\mathrm{g}(\mathbf{x}_j) - \mathbf{y}_j\|^2 + \lambda \|\mathcal{L}\mathrm{g}\|^2$, with some bounded linear operator $\mathcal{L}$, and a positive trade-off parameter λ. Based on the theory of a reproducing kernel Hilbert space, g can be solved as follows([34], chapter 1). Express this mapping function in terms of basis functions $\{\mathrm{U}(\mathbf{x}_i, \mathbf{x}) : i\}$: $\mathrm{g}(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i \mathrm{U}(\mathbf{x}_i, \mathbf{x})$ with vectors α_i to be determined. Select the inner product to be $< \mathrm{U}(\mathbf{s}, \cdot), \mathrm{U}(\mathbf{t}, \cdot) >= \int \mathcal{L}\mathrm{U}(\mathbf{s}, \mathbf{u}) \mathcal{L}\mathrm{U}(\mathbf{t}, \mathbf{u}) d\mathbf{u} = \mathrm{U}(\mathbf{s}, \mathbf{t})$. Then the cost function E_{spline} becomes

$$\sum_{j=1}^{n} \| \sum_{i=1}^{n} \mathrm{U}(\mathbf{x}_i, \mathbf{x}_j) \alpha_i - \mathbf{y}_j \|^2 + \lambda \sum_{i=1, j=1}^{n,n} (\alpha_i \cdot \alpha_j \mathrm{U}(\mathbf{x}_i, \mathbf{x}_j)). \tag{10}$$

Finding $\{\alpha_i\}$ is a least square problem, which can be solved by matrix decompositions SVD or QR.

Different functions U give different splines. In the *thin-plate spline*(TPS), $\mathrm{g}(\mathbf{x})$ is given by $\sum_{i=1}^{n} \alpha_i \mathrm{U}(\mathbf{x}_i, \mathbf{x}) +$ an affine transformation, with $\mathrm{U}(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|^2 \ln \|\mathbf{x} - \mathbf{y}\|$ or $\mathrm{U}(\mathbf{x}, \mathbf{y}) = 1/\|\mathbf{x} - \mathbf{y}\|$, if $N = 2$, or $N = 3$ respectively[3]. If $\mathrm{U}(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2/\sigma^2)$ is used, the spline is called the *radial basis function*(RBF) spline.

Integrating the HD model with splines, we get a non-rigid matching model, whose objective function is : $\min_{\mathrm{g}} \min_{\mathbf{M}^+, \mathbf{M}^-} \{\mathrm{E}_{SHD}(\mathrm{g}, \mathbf{M}^+, \mathbf{M}^-) :=$

$$- \sum_{i=1, j=1}^{n,n} \sqrt{\mu_{i,j}^+ \mu_{i,j}^-} \mathrm{K}(\mathrm{g}(\mathbf{x}_i), \mathbf{y}_j) + \lambda \|\mathcal{L}\mathrm{g}\|^2\}. \tag{11}$$

This optimization problem could be solved by a similar alternating minimization algorithm.

$$\min_{\{r_{i,j}\}, \mathrm{g}} \min_{\mathbf{M}^+, \mathbf{M}^-} \{\bar{E}_{SHD} := \sum_{i=1, j=1}^{n,n} \sqrt{\mu_{i,j}^+ \mu_{i,j}^-}$$

$$(r_{i,j} \log r_{i,j} - r_{i,j} + r_{i,j} \frac{\|\mathrm{g}(\mathbf{x}_i) - \mathbf{y}_j\|^2}{\sigma^2}) + \lambda \|\mathcal{L}\mathrm{g}\|^2\}. \tag{12}$$

Then g is the minimizer of a least square problem with $\{r_{i,j}\}, \mathbf{M}^+, \mathbf{M}^-\}$ fixed. Thus, the overall algorithm is similar to the affine case, except replacing the weighted least square problem in the step 1 by Eq. (10), and replacing $\mathbf{A}\mathbf{x}_i + \mathbf{t}$ by $\mathrm{g}(\mathbf{x}_i)$.

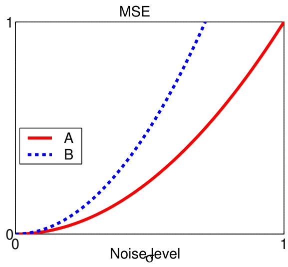

Fig. 4 Illustration of LMSE analysis. Mean square errors are functions of noise level, $(\mathrm{MSE}(\bar{\sigma}))$. When $MSE(0) = 0$, and $d\mathrm{MSE}/d\bar{\sigma}(0) = 0$, then LMSE is related to the second derivatives at $\bar{\sigma} = 0$. MSE of Method A which has a smaller second derivative at 0 is expected to be smaller than that of Method B.

4 Performance comparison between the KC and HD models

To clarify the performances of the KC method and the HD method, we shall provide some quantitative comparisons from these viewpoints separately:

- (i) the mean square error analysis,
- (ii)the robustness to outliers, and
- (iii)the ability to reach the global solution.

In this section, we only discuss (i) and (ii). About (iii), we do not have any theoretical analysis, so the discussion is left in the section 6. Our conclusion is that except (iii), the HD model is expected to outperform the KC model in terms of mean square errors and average shifts.

4.1 MSE analysis

A natural criterion for optimal estimators is the mean square error(MSE)[18], which measures the average squared deviation of the estimator from the true value. Here we shall examine their mean square errors under pure translations, pure rotations and affine transformations. The result shows that the HD method outperforms the KC method when the point-to-point correspondence is identified correctly.

In the following analysis, we only focus on MSE, thus we assume the global solution can be reached without difficulties.

4.1.1 Translations

Given point sets $X := \{\mathbf{x}_i : i = 1, \ldots, n\}$, and $Y := \{\mathbf{y}_i : i = 1, \ldots, n\}$, and $\mathbf{y}_i = \mathbf{x}_i + \bar{\sigma}\eta_i + \mathbf{t}$, where $t \in \mathbb{R}^{N \times 1}$ is a translation vector, and $\{\eta_i : i\}$ are independently identical random vectors with $Normal(0, 1)$. For simplicity, we may assume the ground truth $\mathbf{t} = \mathbf{0}$, i.e., the points are constructed by the rule $\mathbf{y}_i = \mathbf{x}_i + \bar{\sigma}\eta_i$. Let $\mathbf{t}_{HD}$, $\mathbf{t}_{KC}$ be the estimation values in the HD and KC models respectively. Hence $\mathbf{t}$ is a function of $\bar{\sigma}$, and $\mathbf{t} = \mathbf{0}$ at $\bar{\sigma} = 0$.

Define *local mean square error* of $\mathbf{t}$,

$$\mathrm{LMSE}(\mathbf{t}) = \lim_{\bar{\sigma}\to 0} \mathcal{E}[\|\mathbf{t}/\bar{\sigma}\|^2],$$

i.e, this quantity measures the acceleration of MSE of t, when the variance of noise is zero(see Fig. 4). In the following analysis, we shall show

$$\mathrm{LMSE}(\mathbf{t}_{KC}) \geq \mathrm{LMSE}(\mathbf{t}_{HD}).$$

Note that $\lim_{\bar{\sigma}\to 0} \mathbf{y}_i = \mathbf{x}_i$, and $\lim_{\bar{\sigma}\to 0} \mathbf{t} = \mathbf{0}$, then

$$\lim_{\bar{\sigma}\to 0} \exp(-\|\mathbf{y}_j - \mathbf{x}_i - \mathbf{t}\|^2/\sigma^2) = \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2). \tag{13}$$

HD model: As $\bar{\sigma}$ is close to zero enough, we have the closest point property, then the matching links are $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \dots, n\}$. Thus,

$$\mathbf{t}_{HD} = arg\max_{\mathbf{t}} \sum_{i=1}^{n} \exp(-\|\mathbf{y}_i - \mathbf{x}_i - \mathbf{t}\|^2/\sigma^2),$$

or $\mathbf{t}_{HD}$ satisfies

$$\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{x}_i - \mathbf{t}_{HD}) \exp(-\|\mathbf{y}_i - \mathbf{x}_i - \mathbf{t}_{HD}\|^2/\sigma^2) = 0.$$

By Eq. (13), then

$$\begin{aligned}
&\lim_{\bar{\sigma}\to 0} \frac{\mathbf{t}_{HD}}{\bar{\sigma}} \\
&= \lim_{\bar{\sigma}\to 0} \frac{\sum_{i=1}^{n} (\mathbf{y}_i - \mathbf{x}_i) \exp(-\|\mathbf{y}_i - \mathbf{x}_i - \mathbf{t}_{HD}\|^2/\sigma^2)}{\bar{\sigma} \sum_{i=1}^{n} \exp(-\|\mathbf{y}_i - \mathbf{x}_i - \mathbf{t}_{HD}\|^2/\sigma^2)} \\
&= \frac{\sum_{i=1}^{n} \eta_i \exp(-\|\mathbf{x}_i - \mathbf{x}_i\|^2/\sigma^2)}{\sum_{i=1}^{n} \exp(-\|\mathbf{x}_i - \mathbf{x}_i\|^2/\sigma^2)} = \sum_{i=1}^{n} \frac{\eta_i}{n},
\end{aligned}$$

thus $\mathrm{LMSE}(\mathbf{t}_{HD}) =$

$$\lim_{\bar{\sigma}\to 0} \mathcal{E}[\|(\mathbf{t}_{HD}/\bar{\sigma})\|^2] = \sum_{i=1}^{n} \mathfrak{Var}(\eta_i)/n^2 = 1/n.$$

KC model:

$$\mathbf{t}_{KC} = arg\max_{\mathbf{t}} \sum_{i=1,j=1}^{n,n} \exp(-\|\mathbf{y}_j - \mathbf{x}_i - \mathbf{t}\|^2/\sigma^2), i.e.,$$

$$\sum_{i=1,j=1}^{n,n} (\mathbf{y}_j - \mathbf{x}_i - \mathbf{t}_{KC}) \exp(-\|\mathbf{y}_j - \mathbf{x}_i - \mathbf{t}_{KC}\|^2/\sigma^2) = 0.$$

By Eq. (13), $\lim_{\bar{\sigma}\to 0} \frac{\mathbf{t}_{KC}}{\bar{\sigma}}$

$$\begin{aligned}
&= \frac{\sum_{i=1,j=1}^{n,n} (\mathbf{x}_j + \bar{\sigma}\eta_j - \mathbf{x}_i) \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2)}{\bar{\sigma} \sum_{i=1,j=1}^{n,n} \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2)}. \\
&= \frac{\sum_{i=1,j=1}^{n,n} \eta_j \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2)}{\sum_{i=1,j=1}^{n,n} \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2)}.
\end{aligned}$$

By Cauchy-Schwartz inequalities, we have

$$\begin{aligned}
&\mathrm{LMSE}(\mathbf{t}_{KC}) \\
&= \lim_{\bar{\sigma}\to 0} \frac{\sum_{j=1}^{n} \mathfrak{Var}(\eta_j) (\sum_{i=1}^{n} \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2))^2}{(\sum_{i=1,j=1}^{n,n} \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2))^2} \geq \frac{1}{n},
\end{aligned}$$

where equality holds only if $\sum_{j=1}^{n} \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2)$ is a constant independent of i. Therefore, $\mathrm{LMSE}(\mathbf{t}_{KC}) \geq \mathrm{LMSE}(\mathbf{t}_{HD})$.

4.1.2 Local MSE in affine transformations

Given point sets $X := \{\mathbf{x}_i : i = 1, \dots, n\}$, and $Y := \{\mathbf{y}_i : i = 1, \dots, n\}$, and $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \bar{\sigma}\eta_i$, where $\mathbf{A} \in \mathbb{R}^{N\times N}$ is a matrix, and $\{\eta_i : i\}$ are independently identical random vector with $Normal(0, 1)$. [6] Without loss of generality, assume the ground truth $\mathbf{A} = \mathbf{I}$, i.e., the points in Y are constructed by the rule $\mathbf{y}_i = \mathbf{x}_i + \bar{\sigma}\eta_i$(Otherwise let $\mathbf{x}_i \leftarrow \mathbf{A}\mathbf{x}_i$). Let $\mathbf{A}_{HD}$, $\mathbf{A}_{KC}$ be the estimators obtained in the HD and KC models respectively. Hence $\mathbf{A}$ is a function of $\bar{\sigma}$, and $\mathbf{A} = \mathbf{I}$ at $\bar{\sigma} = 0$. In the following analysis, we consider the *Frobenius norm*([30]) $\|\mathbf{B}\|_F^2 = \mathfrak{Tr}(\mathbf{B}^\top\mathbf{B})$, and let the local mean square error of $\mathbf{A}$ be

$$\mathrm{LMSE}(\mathbf{A}) := \sum_{i=1,j=1}^{n,n} \mathcal{E}[\mathbf{B}_{i,j}^2] = \mathbf{E}[\|\mathbf{B}\|_F^2],$$

where $\mathbf{B}$ is the limit matrix $\lim_{\bar{\sigma}\to 0}(\mathbf{A} - \mathbf{I})/\bar{\sigma}$. We shall prove

$$\mathrm{LMSE}(\mathbf{A}_{HD}) \leq \mathrm{LMSE}(\mathbf{A}_{KC}).$$

Note that $\lim_{\bar{\sigma}\to 0} \mathbf{y}_i = \mathbf{x}_i$, then

$$\lim_{\bar{\sigma}\to 0} \{r_{i,j} := \exp(-\|\mathbf{y}_j - \mathbf{A}\mathbf{x}_i\|^2/\sigma^2)\} = \exp(-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2) \tag{14}$$

HD model: As $\bar{\sigma}$ is close to zero enough, by the closest point property, the matching links are $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \dots, n\}$. Note

$$\mathbf{A}_{HD} = (\sum_{i=1,j=1}^{n,n} \mathbf{y}_j\mathbf{x}_i^\top r_{i,j})(\sum_{i=1,j=1}^{n,n} \mathbf{x}_i\mathbf{x}_i^\top r_{i,j})^\dagger.$$

By Eq. (14), then let

$$\begin{aligned}
&\mathbf{B}_{HD} := \lim_{\bar{\sigma}\to 0} \frac{\mathbf{A}_{HD} - \mathbf{I}}{\bar{\sigma}} \\
&= \lim_{\bar{\sigma}\to 0} \frac{(\sum_{i=1}^{n} \mathbf{y}_i\mathbf{x}_i^\top - \mathbf{x}_i\mathbf{x}_i^\top)(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top)^\dagger}{\bar{\sigma}} \\
&= (\sum_{i=1}^{n} \eta_i\mathbf{x}_i^\top)(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^\top)^\dagger,
\end{aligned}$$

[6] This is the case the affine transformation without translations. In general cases, the analysis still works: Since one may consider $\widetilde{\mathbf{x}}_i = [\mathbf{x}_i; 1] \in \mathbb{R}^{(N+1)\times 1}$, and $\widetilde{\mathbf{A}} = [\mathbf{A}, \mathbf{t}] \in \mathbb{R}^{N\times(N+1)}$, then the case $\{\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{t} + \eta_i : i\}$ can be reduced to $\{\mathbf{y}_i = \widetilde{\mathbf{A}}\widetilde{\mathbf{x}}_i + \eta_i : i\}$.

thus

$$\mathrm{LMSE}(A_{HD}) = \mathbf{E}\,[\mathfrak{Tr}(\mathbf{B}_{HD}^{\top}\mathbf{B}_{HD})]$$
$$= \mathfrak{Tr}\left((\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\top})^{\dagger\top}(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\top})(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\top})^{\dagger} \right)^{7}$$
$$= \mathfrak{Tr}((\sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i^{\top})^{\dagger}) = \mathfrak{Tr}((\mathbf{X}\mathbf{X}^{\top})^{\dagger}) = \|\mathbf{X}^{\dagger}\|_F^2,$$

here $\mathbf{X} := [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{N\times n}$.

KC model: From Remark 2, $\mathbf{A}_{KC} = arg\max_{\mathbf{A}}$

$$\sum_{i=1,j=1}^{n,n} (2\exp(-\frac{\|\mathbf{y}_j - \mathbf{A}\mathbf{x}_i\|^2}{\sigma^2}) - \exp(-\frac{\|\mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\|^2}{\sigma^2})).$$
$$\Rightarrow \sum_{i=1,j=1}^{n,n} (\mathbf{A}_{KC}\mathbf{x}_i\mathbf{x}_j^{\top} - \mathbf{y}_j\mathbf{x}_i^{\top})r_{i,j} = 0,$$
$$\Rightarrow \mathbf{A}_{KC} = (\sum_{i=1,j=1}^{n,n} \mathbf{y}_j\mathbf{x}_i^{\top}r_{i,j})(\sum_{i=1,j=1}^{n,n} \mathbf{x}_j\mathbf{x}_i^{\top}r_{i,j})^{\dagger}.$$

where $r_{i,j} = \exp(-\|\mathbf{y}_j - \mathbf{A}\mathbf{x}_i\|^2/\sigma^2)$.

By Eq. (14), let $\mathbf{z}_j = \sum_{i=1}^{n} \mathbf{x}_i r_{i,j}$, then $\mathbf{B}_{KC}$

$$:= \lim_{\bar\sigma\to 0} \frac{\mathbf{A}_{KC} - \mathbf{I}}{\bar\sigma}$$
$$= \lim_{\bar\sigma\to 0} \frac{(\sum_{i=1,j=1}^{n,n}(\mathbf{y}_j - \mathbf{x}_j)\mathbf{x}_i^{\top}r_{i,j})(\sum_{i=1,j=1}^{n,n}\mathbf{x}_j\mathbf{x}_i^{\top}r_{i,j})^{\dagger}}{\bar\sigma}$$
$$= (\sum_{j=1}^{n} \eta_j\mathbf{z}_j^{\top})(\sum_{j=1}^{n} \mathbf{x}_j\mathbf{z}_j^{\top})^{\dagger}.$$

Finally, let $\mathbf{Z} := [\mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n] \in \mathbb{R}^{N\times n}$, then

$$\mathrm{LMSE}(\mathbf{A}_{KC}) = \mathbf{E}\,[\mathfrak{Tr}(\mathbf{B}_{KC}^{\top}\mathbf{B}_{KC})]$$
$$= \mathfrak{Tr}\left((\sum_{i=1}^{n} \mathbf{x}_i\mathbf{z}_i^{\top})^{\dagger\top}(\sum_{i=1}^{n} \mathbf{z}_i\mathbf{z}_i^{\top})(\sum_{i=1}^{n} \mathbf{x}_i\mathbf{z}_i^{\top})^{\dagger} \right)$$
$$= \mathfrak{Tr}((\mathbf{X}\mathbf{Z}^{\top})^{\dagger\top}\mathbf{Z}\mathbf{Z}^{\top}(\mathbf{X}\mathbf{Z}^{\top})^{\dagger}) = \|\mathbf{Z}^{\top}(\mathbf{X}\mathbf{Z}^{\top})^{\dagger}\|_F^2$$
$$\geq \|(\mathbf{X}^{\dagger}\mathbf{X})Z^{\top}(\mathbf{X}\mathbf{Z}^{\top})^{\dagger}\|_F^2 = \|\mathbf{X}^{\dagger}\|_F^2.$$

Here the inequality comes from that the inserted part $\mathbf{X}^{\dagger}\mathbf{X}$ is a projection matrix.[8] Hence, $\mathrm{LMSE}(\mathbf{A}_{KC}) \geq \mathrm{LMSE}(\mathbf{A}_{HD})$.

4.1.3 Local MSE in rotations

Given point sets $X := \{\mathbf{x}_i : i = 1, \ldots, n\}$, and $Y := \{\mathbf{y}_i : i = 1, \ldots, n\}$, and $\mathbf{y}_i = \mathbf{Q}\mathbf{x}_i + \bar\sigma\eta_i$, where $\mathbf{Q} \in \mathbb{R}^{N\times N}$ is an orthogonal matrix, and $\{\eta_i : i\}$ are independently identical random vectors with $Normal(0, 1)$. Without loss of generality, assume the ground truth $\mathbf{Q} = \mathbf{I}$, i.e., the points are constructed by the rule $\mathbf{y}_i = \mathbf{x}_i + \bar\sigma\eta_i$, otherwise let $\mathbf{x}_i$ be $\mathbf{Q}\mathbf{x}_i$. Let $\mathbf{Q}_{HD}$, $\mathbf{Q}_{KC}$ be the estimators obtained in the HD and KC models respectively. Hence $\mathbf{Q}$ is a function of $\bar\sigma$, and $\mathbf{Q} = \mathbf{I}$ at $\bar\sigma = 0$.

Here we still use the Frobenius norm in the local mean square error,

$$\mathrm{LMSE}(\mathbf{Q}) = \sum_{i=1,j=1}^{n,n} \mathcal{E}[\dot{\mathbf{Q}}_{i,j}^2] = \mathbf{E}\,[\|\dot{\mathbf{Q}}\|_F^2],$$

where $\dot{\mathbf{Q}}$ is the limit matrix

$$\lim_{\bar\sigma\to 0} \frac{\mathbf{Q} - \mathbf{I}}{\bar\sigma} = \frac{d\mathbf{Q}}{d\bar\sigma}|_{\bar\sigma=0}.$$

We shall prove that $\mathrm{LMSE}(\mathbf{Q}_{KC}) \geq \mathrm{LMSE}(\mathbf{Q}_{HD})$. Hence, the HD method is expected to have a smaller mean square error in $\mathbf{Q}$.

The following analysis shall involve the second derivatives $\ddot{\mathbf{Q}} := \frac{d^2\mathbf{Q}}{d\bar\sigma^2}|_{\bar\sigma=0}$, we need to handle the computation carefully. Let $r_{i,j} = \exp(\|\mathbf{y}_j - \mathbf{x}_i\|^2/\sigma^2)$(a function of $\bar\sigma$), and

$$\bar r_{i,j} = \lim_{\bar\sigma\to 0} r_{i,j} = \exp(\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2).$$

We need the following fact about the derivatives of $\mathbf{Q}$. From $\mathbf{Q}^{\top}\mathbf{Q} = I$, then by taking derivatives we have $\mathbf{Q}^{\top}\frac{d\mathbf{Q}}{d\bar\sigma} + \frac{d\mathbf{Q}^{\top}}{d\bar\sigma}\mathbf{Q} = 0$, and

$$\mathbf{Q}^{\top}\frac{d^2\mathbf{Q}}{d\bar\sigma^2} + 2\frac{d\mathbf{Q}^{\top}}{d\bar\sigma}\frac{d\mathbf{Q}}{d\bar\sigma} + \frac{d^2\mathbf{Q}^{\top}}{d\bar\sigma^2}\mathbf{Q} = 0.$$

Then at $\bar\sigma = 0$, we have

$$\mathbf{Q}^{\top}\frac{d\mathbf{Q}}{d\bar\sigma} + \frac{d\mathbf{Q}^{\top}}{d\bar\sigma}\mathbf{Q} = 0, \tag{15}$$

and

$$2\dot{\mathbf{Q}}^{\top}\dot{\mathbf{Q}} + (\ddot{\mathbf{Q}} + \ddot{\mathbf{Q}}^{\top}) = 0. \tag{16}$$

Note that $\mathbf{Q}$ minimizes $\sum_{i=1,j=1}^{n,n}\exp(-\|\mathbf{y}_j - Q\mathbf{x}_i\|^2/\sigma^2)$, then taking derivatives and using Eq. (15), $\mathbf{Q}$ satisfies the equation:

$$F(\bar\sigma) := \sum_{i=1,j=1}^{n,n} (\mathbf{x}_i^{\top}\mathbf{Q}^{\top} - \mathbf{y}_j^{\top})\frac{d\mathbf{Q}}{d\bar\sigma}\mathbf{x}_i r_{i,j}$$
$$= -\sum_{i=1,j=1}^{n,n} \mathbf{y}_j^{\top}\frac{d\mathbf{Q}}{d\bar\sigma}\mathbf{x}_i r_{i,j} = 0,$$

then

$$F(0) = -\sum_{i=1,j=1}^{n,n} \mathbf{x}_j^{\top}\dot{\mathbf{Q}}\mathbf{x}_i\bar r_{i,j} = 0$$

[8] For a matrix $\mathbf{R}$, and a projection matrix $\mathbf{P}$, we have

$$\|\mathbf{R}\|_F^2 = \mathfrak{Tr}(\mathbf{R}^{\top}\mathbf{P}^{\top}\mathbf{P}\mathbf{R}) + \mathfrak{Tr}(\mathbf{R}^{\top}(\mathbf{I} - \mathbf{P}^{\top}\mathbf{P})\mathbf{R}) \geq \|\mathbf{P}\mathbf{R}\|_F^2,$$

where of course the number of rows in $\mathbf{R}$ is the same as the number of columns in $\mathbf{P}$.

and

$$
\begin{aligned}
&F'(0) = 0\\
\Rightarrow &\lim_{\bar{\sigma}\to 0} \frac{F(\bar{\sigma}) - F(0)}{\bar{\sigma}} = 0,\\
\Rightarrow &\lim_{\bar{\sigma}\to 0} \frac{1}{\bar{\sigma}} \sum_{i=1,j=1}^{n,n} \left((\mathbf{x}_j + \bar{\sigma}\eta_j)^\top \frac{d\mathbf{Q}}{d\bar{\sigma}} \mathbf{x}_i r_{i,j} - \mathbf{x}_j^\top \dot{\mathbf{Q}} \mathbf{x}_i \bar{r}_{i,j} \right) = 0\\
\Rightarrow &\sum_{i=1,j=1}^{n,n} \left(\mathbf{x}_j^\top \ddot{\mathbf{Q}} \mathbf{x}_i \bar{r}_{i,j} + \eta_j^\top \dot{\mathbf{Q}} \mathbf{x}_i \bar{r}_{i,j} \right) = 0.
\end{aligned}
$$

From Eq. (16), we know that $\dot{\mathbf{Q}}$ must satisfy the equation

$$\sum_{i=1,j=1}^{n,n} \left(-(\dot{\mathbf{Q}}\mathbf{x}_j)^\top \dot{\mathbf{Q}} \mathbf{x}_i \bar{r}_{i,j} + \eta_j^\top \dot{\mathbf{Q}} \mathbf{x}_i \bar{r}_{i,j} \right) = 0. \tag{17}$$

Since we are interested in $\mathrm{LMSE}(\mathbf{Q}) = \mathcal{E}[\|\dot{\mathbf{Q}}\|_F^2]$, we need to get the norm of $\dot{\mathbf{Q}}$. Let $\theta = \|\dot{\mathbf{Q}}\|_F$, and $\hat{\mathbf{Q}} = \dot{\mathbf{Q}}/\theta$, then we have

$$\theta = \frac{\sum_{i=1,j=1}^{n,n} \eta_j^\top \hat{\mathbf{Q}} \mathbf{x}_i \bar{r}_{i,j}}{\sum_{i=1,j=1}^{n,n} (\hat{\mathbf{Q}}\mathbf{x}_j)^\top \hat{\mathbf{Q}} \mathbf{x}_i \bar{r}_{i,j}},$$

and $\mathrm{LMSE}(\mathbf{Q}) = \mathcal{E}[\theta^2]$.

HD model: By the closest point property, we have $\bar{r}_{i,j} = \delta_{i,j}$, then

$$\theta_{HD} = \frac{\sum_{i=1}^n \eta_i^\top \hat{\mathbf{Q}} \mathbf{x}_i}{\sum_{i=1}^n (\hat{\mathbf{Q}}\mathbf{x}_i)^\top \hat{\mathbf{Q}} \mathbf{x}_i}.$$

and then

$$\mathcal{E}[\theta_{HD}^2] = \frac{c \sum_{i=1}^n \|\hat{\mathbf{Q}}\mathbf{x}_i\|^2}{(\sum_{i=1}^n \|\hat{\mathbf{Q}}\mathbf{x}_i\|^2)^2} = \frac{c}{\sum_{i=1}^n \|\hat{\mathbf{Q}}\mathbf{x}_i\|^2},$$

here

$$c_1 := \mathfrak{Var}(\eta_i^\top \frac{\hat{\mathbf{Q}}\mathbf{x}_i}{\|\hat{\mathbf{Q}}\mathbf{x}_i\|}) \tag{18}$$

being a constant, independent of $\hat{\mathbf{Q}}_i$.

KC model: Let $\mathbf{z}_j = \sum_{i=1}^n \bar{r}_{i,j}\mathbf{x}_i$, then

$$\theta_{KC} = \frac{\sum_{j=1}^n \eta_j^\top \hat{\mathbf{Q}} \mathbf{z}_j}{\sum_{j=1}^n (\hat{\mathbf{Q}}\mathbf{x}_j)^\top \hat{\mathbf{Q}} \mathbf{z}_j}.$$

and then

$$\mathcal{E}[\theta_{KC}^2] = \frac{c_1 \sum_{j=1}^n \|\hat{\mathbf{Q}}\mathbf{z}_j\|^2}{(\sum_{j=1}^n (\hat{\mathbf{Q}}\mathbf{x}_j)^\top \hat{\mathbf{Q}}\mathbf{z}_j)^2} \geq \frac{c_1}{\sum_{j=1}^n \|\hat{\mathbf{Q}}\mathbf{x}_j\|^2},$$

where we used Cauchy-Schwartz inequalities and c_1 is the same constant in Eq. (18). Hence, $\mathcal{E}[\theta_{KC}^2] \geq \mathcal{E}[\theta_{HD}^2]$, then $\mathrm{LMSE}(\mathbf{Q}_{KC}) \geq \mathrm{LMSE}(\mathbf{Q}_{HD})$.

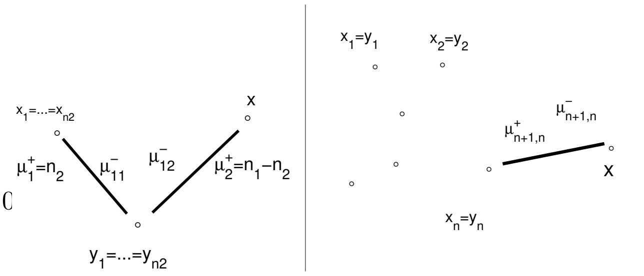

Fig. 5 Illustration of the robustness analysis: the warmup example(Left), and the general case(Right).

4.2 Robustness against outliers

To analyze the robustness under the existence of outliers, we shall compute the average shift(the difference between $\mathbf{t}$ and its ground truth) of the translation estimator $\mathbf{t}$ in noise-free point sets X, Y. That is, let n_1, n_2 be the cardinality of X, Y, and $n_1 > n_2$. And let $X = \{\mathbf{x}_i : i = 1, \ldots, n_2\} \cup \{\mathbf{x}_i : i = n_2 + 1, \ldots, n_1\}$, and $Y = \{\mathbf{y}_i = \mathbf{x}_i + t : i = 1, \ldots, n_2\}$. For simplicity, we call these points $\{\mathbf{x}_i : i = n_2+1, \ldots, n_1\}$(which have no correspondents in Y) outliers, and the remaining points (which have correspondents in Y) non-outliers.[9] Our goal is to show the following results:

- Under the kernel scale σ fixed, if the point-to-point correspondence can be identified correctly, then the HD method has a stronger robustness than the KC method.
- Consider the bias caused by a genuine outlier that has a far distance to the rest of points. Both methods have similar results if different kernel scales are used, $\sigma_{HD}^2 = \sigma_{KC}$.

A heuristic (warm-up) example:(See the left in Fig. 5.) Assume the ground truth $\mathbf{t} = \mathbf{0}$. Assume that n_2 points $\{\mathbf{x}_i : i = 1, \ldots, n_2\}$ are located at $\mathbf{0}$, and the remaining $n_1 - n_2$ points are located as $\mathbf{x}$; all n_2 points in Y are located at $\mathbf{0}$. Then this is equivalent to having three weighted points, $\{(\mathbf{0}, \mu_1^+ := n_2), (\mathbf{x}, \mu_2^+ := n_1 - n_2)\} \in X$ and $\{(\mathbf{0}, \mu_1^- := n_2)\} \in Y$, where the second component refers to the weight.

HD model: The estimator

$$\mathbf{t}_{HD} = arg \max_{\mathbf{t}} \max_{\mu_{1,1}^- + \mu_{2,1}^- = \mu_1^-}$$

$$\sqrt{\mu_1^+ \mu_{1,1}^-} \exp(-\frac{\|\mathbf{t}\|^2}{\sigma^2}) + \sqrt{\mu_2^+ \mu_{2,1}^-} \exp(-\frac{\|\mathbf{x}+\mathbf{t}\|^2}{\sigma^2}). \tag{19}$$

[9] In statistics, an outlier is an observation that is numerically distance from the rest of data. We shall return to this definition in the end of this subsection.

By Cauchy-Schwartz inequalities , we have

$$\frac{\mu^-_{1,1}}{\mu^-_{2,1}} = \frac{\sqrt{\mu^+_1}\exp(-\frac{\|\mathbf{t}\|^2}{\sigma^2})}{\sqrt{\mu^+_2}\exp(-\frac{\|\mathbf{x}+\mathbf{t}\|^2}{\sigma^2})}.$$

Substituting it into Eq. (19), we have $\mathbf{t}_{HD} =$

$$arg\max_{\mathbf{t}}\left(\mu^+_1\exp(-\frac{2\|\mathbf{t}\|^2}{\sigma^2}) + \mu^+_2\exp(-\frac{2\|\mathbf{x}+\mathbf{t}\|^2}{\sigma^2})\right). \tag{20}$$

Hence, $\mathbf{t} = \mathbf{t}_{HD}$ satisfies

$$\mathbf{t}\exp(-\frac{2\|\mathbf{t}\|^2}{\sigma^2}) + \alpha(\mathbf{t}+\mathbf{x})\exp(-\frac{2\|\mathbf{x}+\mathbf{t}\|^2}{\sigma^2}) = 0, \tag{21}$$

with $\alpha := \mu^+_2/\mu^+_1$.

To study the influence of σ and $\mathbf{x}$ on the estimator t, we use one common tool in quantitative (infinitesimal) robustness[13]: the *influence function*, which is defined by the limit:

$$\text{IF}(\mathbf{x},\sigma) = \lim_{\alpha\to 0}\frac{\text{t}(\alpha,\sigma,\mathbf{x}) - \text{t}(0,\sigma,x)}{\alpha},$$

where $\text{t}(\alpha,\sigma,\mathbf{x}) = \mathbf{t}$ is viewed as a statistical functional, depending on the ratio of $\alpha = \mu^+_2/\mu^+_1$, and also σ, $\mathbf{x}$.

When $\mu^+_2 = 1$, (i.e., one outlier is added), this function describes the infinitesimal estimator shifts under the infinitesimal changes in the point set configuration– by adding one point at $\mathbf{x}$. Here is one interpretation about the ratio of weights α approaching 0: Duplicate a number of copies of non-outliers, then the limit describes the product of the shift in t and the total number of non-outliers ($1/\alpha$) as the number of copies grows to infinity.

To compute the influence function, expanding $\exp(-2\|\mathbf{x}+\mathbf{t}\|^2/\sigma^2)$ in Eq. (21), and canceling the nonzero factor $\exp(-2\|\mathbf{t}\|^2/\sigma^2)$, then we have

$$\frac{\mathbf{t}_{HD}}{\alpha} = -(\mathbf{t}_{HD}+\mathbf{x})\exp(-\frac{2\|\mathbf{x}\|^2 + 4\mathbf{x}\cdot\mathbf{t}_{HD}}{\sigma^2}).$$

Note that $\lim_{\alpha\to 0}\text{t}_{HD}(\alpha,\sigma,\mathbf{x}) = \text{t}_{HD}(0,\sigma,\mathbf{x}) = 0$, then by taking the limits $\alpha \to 0$ on both sides, we have

$$\lim_{\alpha\to 0}\frac{\mathbf{t}_{HD}}{\alpha} = \mathbf{x}\exp(-\frac{2\|\mathbf{x}\|^2}{\sigma^2}).$$

Thus the influence function is given by

$$\text{IF}_{HD}(\mathbf{x},\sigma) = \lim_{\alpha\to 0}\frac{\text{t}(\alpha,\sigma,\mathbf{x}) - \text{t}(0,\sigma,\mathbf{x})}{\alpha} = \mathbf{x}\exp(-\frac{2\|\mathbf{x}\|^2}{\sigma^2}).$$

Hence, the influence function is an increasing function of σ. We conclude, that the smaller the kernel scale σ is, the smaller the shift $\mathbf{t}_{HD}$ is.

KC model: $\mathbf{t}_{KC} = arg\max_{\mathbf{t}}$

$$n_2 n_2\exp(-\frac{\|\mathbf{t}\|^2}{\sigma^2}) + (n_1 - n_2)n_2\exp(-\frac{\|\mathbf{x}+\mathbf{t}\|^2}{\sigma^2}),$$

i.e., $\mathbf{t}_{KC} = arg\max_{\mathbf{t}}$

$$\mu^+_1\exp(-\frac{\|\mathbf{t}\|^2}{\sigma^2}) + \mu^+_2\exp(-\frac{\|\mathbf{x}+\mathbf{t}\|^2}{\sigma^2}). \tag{22}$$

Comparing Eq. (20,22), we have $\mathbf{t}_{HD} = \mathbf{t}_{KC}$ provided that $2\sigma^2_{KC} = \sigma^2_{HD}$.

The general case: Now we consider $X = \{\mathbf{x}_i : i = 1,\ldots,n\}\cup\mathbf{x}$, and $Y = \{\mathbf{y}_i = \mathbf{x}_i + \mathbf{t} : i = 1,\ldots,n\}$ with the ground truth $\mathbf{t} = 0$(see Fig. 5). [10] For the sake of notion simplicity, in the following, let α be the weight of the outlier, instead of the ratio.

KC model: The estimator $\mathbf{t}_{KC} = arg\max_{\mathbf{t}}$

$$\sum_{i=1,j=1}^{n,n}\exp(-\|\mathbf{x}_i - \mathbf{y}_j + \mathbf{t}\|^2/\sigma^2) + \alpha\sum_{j=1}^{n}\exp(-\|\mathbf{x} - \mathbf{y}_j + \mathbf{t}\|^2/\sigma^2),$$

then

$$\sum_{i=1,j=1}^{n,n}\frac{(\mathbf{x}_i - \mathbf{y}_j + \mathbf{t}_{KC})\exp(-\|\mathbf{x}_i - \mathbf{y}_j + \mathbf{t}_{KC}\|^2/\sigma^2)}{\alpha}$$

$$= -\sum_{j=1}^{n}(\mathbf{x} - \mathbf{y}_j + \mathbf{t}_{KC})\exp(-\|\mathbf{x} - \mathbf{y}_j + \mathbf{t}_{KC}\|^2/\sigma^2).$$

Using $\sum_{i=1,j=1}^{n,n}(\mathbf{x}_i - \mathbf{x}_j)\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2) = 0$, the assumption $\mathbf{y}_i = \mathbf{x}_i$ and $\lim_{\alpha\to 0}\text{t}_{KC}(\alpha,\sigma,\mathbf{x}) = 0$, we have

$$\lim_{\alpha\to 0}\frac{\mathbf{t}_{KC}}{\alpha}\left(\sum_{i=1,j=1}^{n,n}\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)\right)$$

$$= -\sum_{j=1}^{n}(\mathbf{x} - \mathbf{x}_j)\exp(-\|\mathbf{x} - \mathbf{x}_j\|^2/\sigma^2).$$

Hence, the influence function is

$$\text{IF}_{KC}(\mathbf{x},\alpha) = \lim_{\alpha\to 0}\frac{\text{t}_{KC}(\alpha,\sigma,\mathbf{x}) - \text{t}_{KC}(0,\sigma,\mathbf{x})}{\alpha}$$
$$= \frac{\sum_{j=1}^{n}(\mathbf{x}_j - \mathbf{x})\exp(-\|\mathbf{x} - \mathbf{x}_j\|^2/\sigma^2)}{\sum_{i=1,j=1}^{n,n}\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)}.$$

HD model:

$$\mathbf{t}_{HD} = arg\max_{\mathbf{t}}\max_{\{\mathbf{M}^+,\mathbf{M}^-\}}$$

$$\sum_{i=1,j=1}^{n,n}\sqrt{\mu^+_{i,j}\mu^-_{i,j}}\exp(-\|\mathbf{x}_i - \mathbf{y}_j + \mathbf{t}\|^2/\sigma^2).$$

[10] Here we only focus on one outlier. The infinitesimal shift of multiple outliers is simply the sum of the infinitesimal shifts caused by each individual outlier.

Suppose the closest point of $\mathbf{x}$ in Y is $\mathbf{y}_n$(see the right in Fig. 5), i.e., $\|\mathbf{y}_i - \mathbf{x}\| > \|\mathbf{y}_n - \mathbf{x}\|$ for $i = 1, \ldots, n-1$. Then adding an outlier $\mathbf{x}$, the set of matching links changes from $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \ldots, n\}$ to $\{(\mathbf{x}_i, \mathbf{y}_i) : i = 1, \ldots, n\} \cup \{(\mathbf{x}, \mathbf{y}_n)\}$. And we can show that the correspondence matrices are described by

$$\mu^+_{i,j} = \delta_{i,j} = \mu^-_{i,j} \text{ for } 1 \le i \le n, 1 \le j \le n-1,$$

$$\text{and } \mu^+_{n+1,j} = 0 = \mu^-_{n+1,j} \text{ for } 1 \le j \le n-1$$

(See the appendix B.4). Then

$$\mathbf{t}_{HD} = arg \max_{\mathbf{t}}((n-1)\exp(-\frac{\|\mathbf{t}\|^2}{\sigma^2}) + \sqrt{\exp(-\frac{2\|\mathbf{t}\|^2}{\sigma^2}) + \alpha \exp(-\frac{2\|\mathbf{x} - \mathbf{y}_n + \mathbf{t}\|^2}{\sigma^2})}),$$

i.e., $\mathbf{t} = \mathbf{t}_{HD}$ satisfies $(n-1)\mathbf{t}\exp(-\frac{\|\mathbf{t}\|^2}{\sigma^2})$

$$= -\frac{\mathbf{t}\exp(-\frac{2\|\mathbf{t}\|^2}{\sigma^2}) + \alpha(\mathbf{x} - \mathbf{y}_n + \mathbf{t})\exp(-\frac{2\|\mathbf{x}-\mathbf{y}_n+\mathbf{t}\|^2}{\sigma^2})}{\sqrt{\exp(-\frac{2\|\mathbf{t}\|^2}{\sigma^2}) + \alpha\exp(-\frac{2\|\mathbf{x}-\mathbf{y}_n+\mathbf{t}\|^2}{\sigma^2})}}. \tag{23}$$

Dividing Eq. (23) by α and taking limits, then

$$\lim_{\alpha\to 0}(n-1)\frac{\mathbf{t}_{HD}}{\alpha} = \lim_{\alpha\to 0} -\frac{\mathbf{t}_{HD}}{\alpha} - (\mathbf{x} - \mathbf{y}_n)\exp(-\frac{2\|\mathbf{x} - \mathbf{y}_n\|^2}{\sigma^2}).$$

Hence the influence function is

$$\mathrm{IF}_{HD}(\mathbf{x}, \sigma) = \lim_{\alpha\to 0} \frac{\mathrm{t}_{HD}(\alpha, \sigma, \mathbf{x}) - \mathrm{t}_{HD}(0, \sigma, \mathbf{x})}{\alpha}$$

$$= \frac{(\mathbf{y}_n - \mathbf{x})}{n}\exp(-\frac{2\|\mathbf{x} - \mathbf{y}_n\|^2}{\sigma^2}).$$

As $\mathbf{x}$ is far away from $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, i.e., a genuine outlier, then these two influence functions $\mathrm{IF}_{HD}, \mathrm{IF}_{KC}$ are comparable only when $\sigma^2_{HD} = 2\sigma^2_{KC}$. In other words, choosing $\sigma_{KC} = \sigma_{HD}$ and x far away from $\{x_1, \ldots, x_n\}$, the HD method has a better robustness against outliers.

5 Numerical experiments

Here we provide three series of experiments(**I,II,III**). There are quite a few existing methods outperforms the ICP method, for instance, EM-ICP[11], RPM[7], and KC methods[31]. In (I), we provide several comparison experiments between the KC method, and our method. In our experiments, the HD method outperforms the KC method.

(I) In (i)(ii)(iii), we use two 1D point-sets $X = \{\mathbf{x}_i = i - 1 : i = 1, \ldots, 40\}$ and $Y = \{\mathbf{y}_i = \mathbf{x}_i + t : i = 1, \ldots, 40\}$, with the ground truth $t = 0.2$. Our goal here is to estimate the translation $\mathbf{t}$.

(i) First, add noises to the point set Y, i.e., $\mathbf{y}_i = \mathbf{x}_i + \mathbf{t} + Normal(0, 0.5^2)$. we match two point-sets under different kernel scales, $\sigma = 2, 4, \ldots, 20$. The average shift/variance of each method is plotted in the left/right of the top row of Fig. 8, based on 40 samples. The HD method has smaller variances.

(ii)Next, add 2 outliers to both X, Y. We compare $\mathbf{t}_{HD}, \mathbf{t}_{KC}$(based on 20 samples). To clarify their performances, we consider two different manners to generate outliers.

- The second row of Fig. 8 shows a case, in which outliers $\mathbf{x}_{41}, \mathbf{x}_{42}$ sampled from $(-5, -4)$ uniformly and y_{41}, y_{42} from $(44, 45)$ uniformly.
- The third row shows a case, in which all four outliers are sampled from $(0, 20)$ uniformly.

These results show that the HD method performs better.

(iii) Next, we compare the HD method with the RPM method. To visualize the performance of RPM method, we plot the results at each fixed values of σ in the fourth row. Similar to (iii), $\mathbf{x}_{41}, \mathbf{x}_{42}$ are sampled from $(0, 1)$ and $\mathbf{y}_{41}, \mathbf{y}_{42}$ from $(30, 31)$. When the kernel scale is small, both methods perform similarly(because the correspondence can be identified correctly).

(iv)We compare their performances in matching 2D/3D point-sets under the existence of outliers.
First, we match two 2D point-sets X, Y under different numbers of outliers. Let X be the point-set(100 points) shown in Fig. 8, and construct Y by a randomly generated rigid transformation, i.e., $Y = \mathbf{Q}X + \mathbf{t}$ where $\mathbf{Q}, \mathbf{t}$ are a rotation with angle θ (sampled in uniformly $(0, \pi/5)$ and a 2D translation $\mathbf{t} = [t_x, t_y]^\top$ in uniformly $(0, 0.1) \times (0, 0.1)$). Different numbers of outliers are sampled from the unit square $(0, 1) \times (0, 1)$ randomly and added to Y. Their average estimator errors of (θ, t_x, t_y) in the l^2 norm (based on 5 samples, and $\sigma^2_{HD} = \sigma^2_{KC} = 0.2$) are listed as follows.

# of outliers	KC	Ours
2	1.00e-3	1.9e-4
4	1.10e-3	2.9e-4
6	7.8e-4	3.9e-4
8	1.55e-3	6.7e-4
10	1.68e-3	1.08e-3
12	2.99e-3	5.9e-4

Secondly, we match two 3D point-sets, X, Y, using different kernel scales. Each point-set comprises 300 points, among which 10 points are outliers generated by $Normal([0, 0, 0]^\top, 3)$. In X, the first 290 points $\mathbf{x}_i = [.25\cos(i\pi/3), \sin(i\pi/3), i/30]^\top$ for $i = 1, \ldots, 290$; in Y, $\mathbf{y}_i = \mathbf{Q}\mathbf{x}_i$ for $i = 1, \ldots, 290$ where $\mathbf{Q}$ is a 3×3 orthogonal matrix

$\begin{pmatrix} .86, & .20, & .47 \\ -.17, & .98, & -.10 \\ -.15, & .00, & .88 \end{pmatrix}$. Then the estimator error of $\mathbf{Q}$ are listed as follows (based on 5 samples).

σ^2	KC	Ours
.04	2.4e-4	1.4e-4
.08	7.5e-4	5.5e-4
.16	19.4e-4	15.5e-4
.32	40.5e-4	34.2e-4

From these simple experiments, we can see clearly the advantage of the HD method: Through identifying the correspondence or outliers (provided that the unknown correspondence exists) the HD method has smaller mean-square-errors and smaller average shifts than the KC method. These results are consistent with the mathematical analysis in the previous section.

Last, the efficiency of the proposed Algorithm 1. can be seen as follows. Consider two algorithms(Alg. 1. v.s. matlab toolbox) to implement the KC approach in 2D experiment (iv). The speed of our method(Alg. 1) is much faster than the speed of the gradient methods in the matlab optimization toolbox (suggested in [14]) based on same relative errors 10^{-6}(both matlab codes).

	Alg. 1.	Matlab toolbox
time(sec)	0.6	9.9
# of iteration	30	134

Next, experiments **II** and **III** show some non-rigid matching results. Comparisons with other state-of-art methods are beyond the scope of this paper, which are left in future works.

(II) We match two objects using the non-rigid algorithm in section 3.3. RBF splines are used here. First, we match two point-sets with unit weights $\mu_i^+ = \mu_j^- = 1$. The matching result is shown in Fig. 6. Then we match two sets of endocardial/epicardial boundaries in two phases, ED and ES. In the ED phase, $50, 25$ points are sampled on endocardial/epicardial boundaries respectively(let μ_i^+ be 0.5 and 1 respectively). In the ES phase, $25, 50$ points are sampled on epicardial/endocardial boundaries(letting μ_j^- be 1 and 0.5 respectively). The matching result is shown in Fig.7.

(III) Here is an experiment related to the application of the distance function (4) in the multiple-object registration. The similarity function E_{HD} can be applied to match a large set of objects(point-sets or curves) simultaneously, such that the missing correspondence information is rebuilt and non-rigid transformations among objects have optimal low dimensional representations. More precisely, given m objects $\mathcal{O}_1, \ldots, \mathcal{O}_m$, the goal is to find an optimal subspace L with low dimension, so that the projected objects $\{\mathcal{P}_i : i = 1, \ldots, m\}$ on L maximize the overall similarities $\sum_{i=1}^m \mathrm{E}_{HD}(\mathcal{O}_i, \mathcal{P}_i)$, see Fig. 9, which can be regarded as a minimization of the sum of distances $(\sum_{i=1}^m \mathrm{M}_{HD}(\mathcal{O}_i, \mathcal{P}_i))$ perpendicular to an optimal hyperplane L.

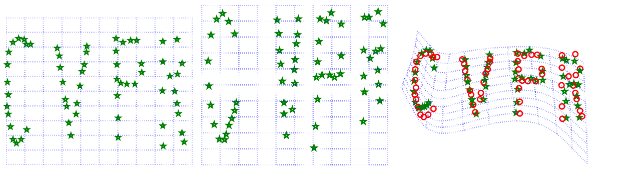

Fig. 6 Matching two point-sets(right). The left and middle sub-figures show the given point-sets. The right sub-figure shows the matching result.

This problem can be solved by repeating these two steps: first, given correspondence matrices, find one mean object and the optimal subspace L with bases $\{\mathbf{v}_1, \mathbf{v}_2, \ldots\}$ using the standard principal component analysis(PCA)[8]; secondly, given the mean object and the optimal subspace, find the projected objects and update the correspondence matrices. In the second steps, we need to solve m correspondence subproblems. [11]

In the experiment, we apply the idea to a large set of point-sets $\{\mathcal{O}_i : i = 1, \ldots, 91\}$, 91 cardiac contours from 91 subjects. All contours are in the ED phase. The assumption here is that these cardiac contours could/might be deformed from a "mean cardiac contour",$\bar{\mathcal{P}}$.

First, align them by proper translations, rotations and scalings. The result is shown in the top left sub-figure in Fig. 10. Then we use the above alternating method to find an optimal two dimensional subspace L spanned by bases $\mathbf{v}_1, \mathbf{v}_2$ to represent deformations. The right sub-figure shows the mean cardiac contour and two singular vectors(the first singular vector in blue thicker line-segments, and the second singular vector in red thinner ones). Then locate all cardiac contours at the position (β_1, β_2) in the bottom sub-figure, where β_1, β_2 are the projection of $\mathcal{P}_i - \bar{\mathcal{P}}$ on bases $\mathbf{v}_1, \mathbf{v}_2$ respectively. The x-axis represents β_1, and the y-axis represents β_2. Obviously a larger value β_1 corresponds to a more slender cardiac contour, and a smaller value β_2 leads to a more symmetric cardiac contour. Finally, all the "residue contours"(removing two principle components $\beta_1\mathbf{v}_1 + \beta_2\mathbf{v}_2$ from $\mathcal{O}_i$) are in the top middle sub-figure.

6 Discussions and conclusion

One issue should be discussed here: How to choose the kernel scale? A proper value σ should be decided based on two criteria: the robustness against outliers and the

[11] This approach (estimating a mean object and correspondences alternatively) in aligning multiple point-sets has been used in many research works, e.g. [35][4].

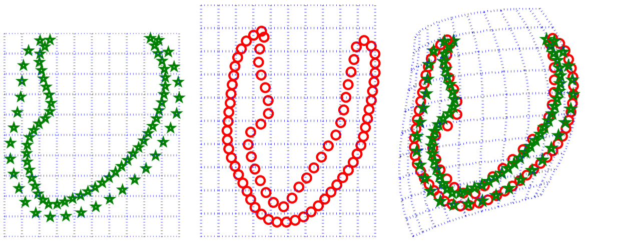

Fig. 7 Matching cardiac contours(right). The left is ED-Endo& ED-Epi. The middle is ES-Endo& ES-Epi.

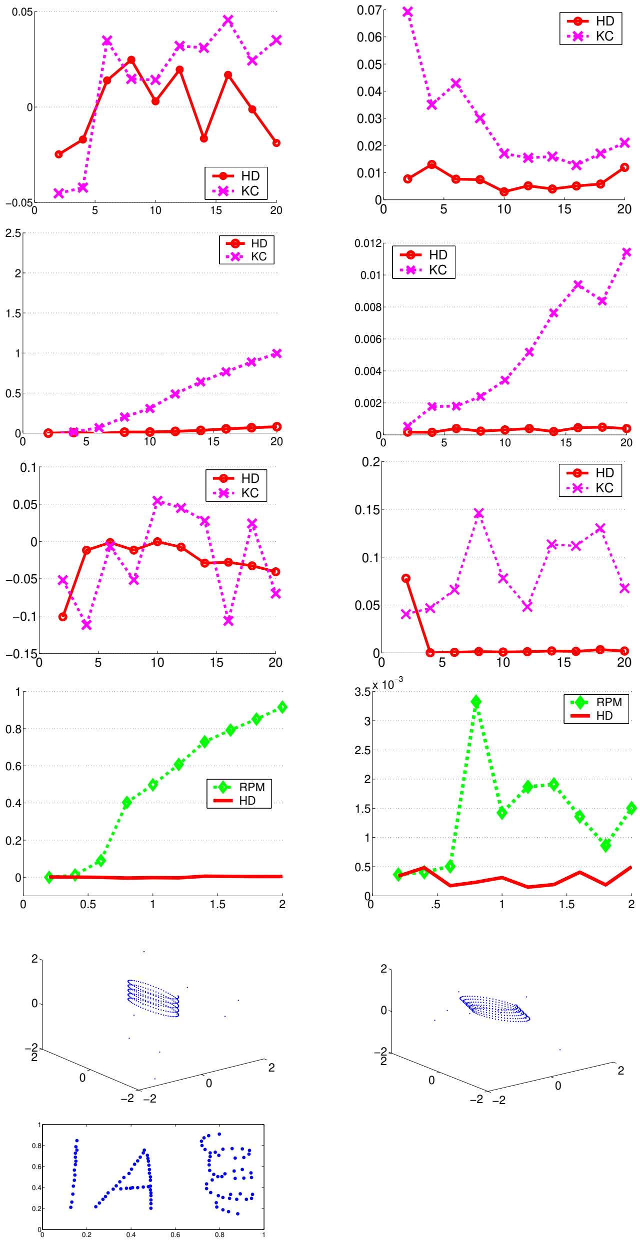

Fig. 8 Four rows display the estimation result in average shift(left) and variance(right) using the HD method, the RPM method and the KC method. Their X-axis represents the kernel scale σ. The fifth row shows the 3D point-sets X(left) and Y(right) and the bottom shows the 2D point-set X for the experiment(iv).

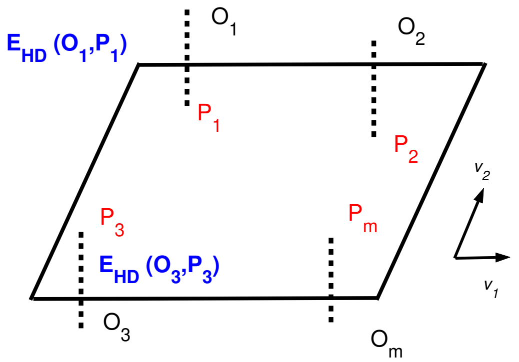

Fig. 9 Illustration of the combination of PCA and the HD method. $\mathcal{P}_i$ is the projection of $\mathcal{O}_i$ on the optimal subspace spanned by $\mathbf{v}_1, \mathbf{v}_2$.

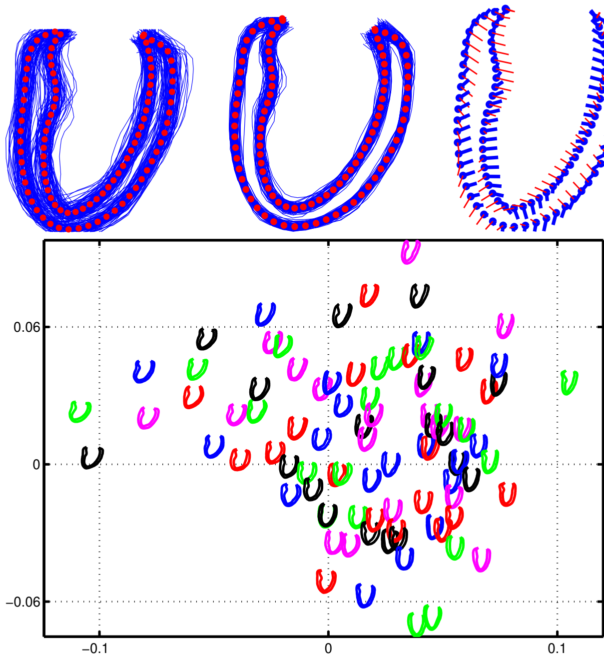

Fig. 10 Matching cardiac contours. The top left sub-figures shows the point-sets– 91 cardiac contours. The red dotted curve is the mean shape. The middle is the "residue contours" given by $\mathcal{O}_i - \beta_1 \mathbf{v}_1 - \beta_2 \mathbf{v}_2$. The right sub-figure shows these 2 basis functions ($\mathbf{v}_1/\mathbf{v}_2$ in blue/red line-segments). The bottom row shows the optimal deformation subspace.

ability to reach the global optimal solution. A smaller value of σ yields the stronger robustness against outliers, but has a higher chance in getting stuck at a non-global local solution.

Empirically both methods performs quite well in matching two point sets, unless the kernel size is chosen too small. Based on this, maybe a good strategy is: start with the KC method with a larger kernel scale, and then proceed with the HD method with a smaller kernel scale.

According to the robustness analysis, through identify correspondences, the *effective* kernel scale σ_{eff} has

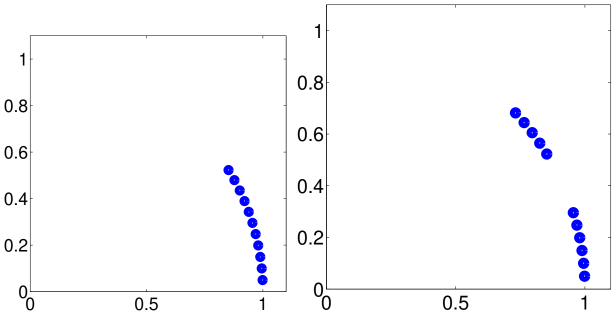

Fig. 11 Illustration of the point set X in example I(Left) and II(Right).

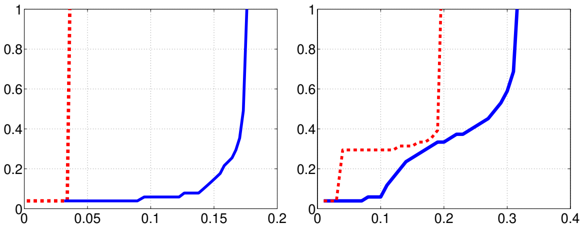

Fig. 12 Graphs of functions $\theta_0(\sigma_{HD})$(in blue, solid curves), and $\theta_0(\sigma_{KC})$(in red, dashed curves). Left: Example I. Right: Example II.

the relation: $\sigma_{eff}^2 = 2\sigma_{KC}^2 = \sigma_{HD}^2$. From this view point, we expect that under the same kernel scale, the HD model has a higher chance in getting stuck at a non-global local solution. Here we provide two numerical examples, in which Y is a rotation of X by angle $\theta \in [0,1]$. We plot the relations between the kernel scale σ_{HD}(or σ_{KC}) and the *convergence range* for the global optimal solution. The convergence range is given by $[0, \theta_0]$, where θ_0 is the maximum rotation angle θ under which the global solution can be reached.

Example I: Let $\{\mathbf{x}_i = [\cos(i/20), \sin(i/20)]^\top : i = 1, \ldots, 11\}$.
Example II: Let $\{\mathbf{x}_i = [\cos(i/20), \sin(i/20)]^\top : i = 1, \ldots, 6\} \cup \{\mathbf{x}_i = [\cos((i+4)/20), \sin((i+4)/20)]^\top : i = 7, \ldots, 11\}$(see Fig. 11).

The results are in Fig. 12. These examples show that both models can reach the global solutions provided that σ_{HD}, σ_{KC} larger enough. And under small kernel scales $\sigma_{HD} = \sigma_{KC}$, the convergence range of the KC method is larger than that of the HD method. Hence the KC model performs better regarding to this issue.

Remark 3 In case of translation estimators with point sets $\{\mathbf{x}_i : i = 1, \ldots, n\}$ and $\{\mathbf{y}_i = \mathbf{x}_i + \mathbf{t} : i = 1, \ldots, n\}$, the global solution can be easily reached (according to section 3.2, let $s_{i,j} = 1$ for that the mass centers of X and Y are matched at the first iteration). Hence, to distinguish their differences, we provide the above rotation examples.

6.1 Conclusion

In this article, we present the HD method to solve point-sets/shape matching problems. Since the HD model bears a resemblance to the KC model, it inherits the robustness from the KC method. The difference lies that the point-to-point correspondence is computed in the HD model. We also derived the associated algorithms. Based on the mean square error and robustness analysis, when the correspondence is correctly identified and the same kernel scale is used, the HD method is expected to outperform the KC method in terms of both average shifts and mean square errors, which is also validated in numerical experiments. Hence, it is fair to say that the HD method provides more accurate results than the KC method.

A Mathematical properties

A.1 Algorithms and optimal conditions

Here we provide some mathematical analysis on E_{HD}.[12] Form the Lagrangian function[1] $\mathrm{L}(\mathbf{M}^+, \mathbf{M}^-) :=$

$$\sum_{i=1,j=1}^{n_1,n_2} 2\sqrt{\mu_{i,j}^+ \mu_{i,j}^-}\mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)$$

$$+\sum_{i=1}^{n_1} a_i(\mu_i^+ - \sum_{j=1}^{n_2} \mu_{i,j}^+) + \sum_{j=1}^{n_2} b_j(\mu_j^- - \sum_{i=1}^{n_1} \mu_{i,j}^-)$$

with multipliers $\{a_i, b_j : i, j\}$. Then the optimal values of $\mathbf{M}^+, \mathbf{M}^-$ satisfies

$$\frac{\partial \mathrm{L}}{\partial \mu_{i,j}^+} \geq 0, \ \frac{\partial \mathrm{L}}{\partial \mu_{i,j}^-} \geq 0.$$

From $\partial \mathrm{L}/\partial \mu_{i,j}^+ \geq 0$, we have $\sqrt{\mu_{i,j}^-/\mu_{i,j}^+}\mathrm{K}(\mathbf{x}_i, \mathbf{y}_j) \leq a_i$ and equality holds when $\mu_{i,j}^+ > 0$. Then

$$\mu_{i,j}^- \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2 = a_i^2 \mu_{i,j}^+ \text{for each } i. \tag{24}$$

To solve for a_i, summing up Eq. (24) over each j, we have

$$\mu_i^+ a_i^2 = \sum_{j=1}^{n_2} \mu_{i,j}^- \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2, \tag{25}$$

$\Rightarrow$ given $\mathbf{M}^-$, then the optimal selection of $\mathbf{M}^+$ is

$$\mu_{i,j}^+ = \frac{\mu_i^+ \mu_{i,j}^- \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2}{\sum_{j=1}^{n_2} \mu_{i,j}^- \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2}. \tag{26}$$

Similarly, from $\frac{\partial \mathrm{L}}{\partial \mu_{i,j}^-} \geq 0$, we have $\sqrt{\mu_{i,j}^+/\mu_{i,j}^-}\mathrm{K}(\mathbf{x}_i, \mathbf{y}_j) \leq b_j$ and

$$\mu_{i,j}^+ \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2 = b_j^2 \mu_{i,j}^-. \tag{27}$$

Then $b_j^2 = \sum_{i=1}^{n_1} \mu_{i,j}^+ \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2/\mu_j^-$,
$\Rightarrow$ given $\mathbf{M}^+$ fixed, the optimal selection of $\mathbf{M}^-$ is

$$\mu_{i,j}^- = \frac{\mu_j^- \mu_{i,j}^+ \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2}{\sum_{i=1}^{n_1} \mu_{i,j}^+ \mathrm{K}(\mathbf{x}_i, \mathbf{y}_j)^2}. \tag{28}$$

[12] The analysis in Appendix A comes from the preprint(the attached file). For the completeness of this paper, we list them here.

A.2 Sparsity of the matching structure

From these two optimal conditions(26,28), we have $\mu^+_{i,j} = 0$ if and only if $\mu^-_{i,j} = 0$. Finally, from Eq. (24, 27), we get

$$\mu^+_{i,j}\mu^-_{i,j}(\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j)^2 - a_i b_j) = 0,$$

which implies, $\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j)^2 = a_i b_j$ for each pair (i,j) with $\mu^+_{i,j}\mu^+_{i,j} > 0$, i.e., $(\mu^+_{i,j}, \mu^-_{i,j})$ is a matching pair. Note that the matrix with entries $\{a_i b_j : i,j\}$ has rank one, but the proximity matrix in general $\{\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j)^2 : i,j\}$ has rank $\min(n_1,n_2)$. So, in case of rank $\min(n_1,n_2)$, the matching pairs are sparse and they do not form any cycle. Since the total number of nodes is n_1+n_2, then the maximal number of matching pairs(edges) is n_1+n_2-1(Proposition 2, page 22,[24]).

A.3 Duality

From the optimal conditions, we have

$$\sqrt{\mu^-_{i,j}\mu^+_{i,j}}\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j) = a_i\mu^+_{i,j} = b_j\mu^-_{i,j},$$

then the Lagrangian function can be reformulated as follows (its dual problem):

$$\min_{\{a_i>0,b_j>0:i,j\}} \sum_{i=1}^{n_1} a_i\mu_i + \sum_{j=1}^{n_2} b_j\mu_j,$$

subject to $a_i b_j \ge \mathrm{K}(\mathbf{x}_i,\mathbf{y}_j)^2$. The duality can be easily seen by the inequality of arithmetic and geometric means :

$$\sum_{i=1,j=1}^{n_1,n_2} 2\sqrt{\mu^+_{i,j}\mu^-_{i,j}}\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j) \le \sum_{i=1,j=1}^{n_1,n_2} 2\sqrt{\mu^+_{i,j}\mu^-_{i,j}a_i b_j}$$

$$\le \sum_{i=1,j=1}^{n_1,n_2}\mu^+_{i,j}a_i + \sum_{i=1,j=1}^{n_1,n_2}\mu^-_{i,j}b_j = \sum_{i=1}^{n_1} a_i\mu_i + \sum_{j=1}^{n_2} b_j\mu_j.$$

This result shows that the minimum of the dual problem is not less than the maximum of the original problem, and in fact, two optimal values are equal, i.e., no duality gap. Since the second equality holds at $a_i\mu^+_{i,j} = b_j\mu^-_{i,j}$ for all i,j, then the optimality of a solution $\{a_i, b_j, \mu^+_{i,j}, \mu^-_{i,j} : i,j\}$ can be verified by checking (i) $a_i\mu^+_{i,j} = b_j\mu^-_{i,j}$ and (ii) $\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j)^2 \le a_i b_j$ and of course (iii) the feasibility of $\{\mu^+_{i,j}, \mu^-_{i,j}\}$: $\sum_j \mu^+_{i,j} = \mu^+_i$, $\sum_i \mu^-_{i,j} = \mu^-_j$.

B Proofs

B.1 Proof of nonnegativity

Since $\mathrm{K}(\mathbf{x},\mathbf{y}) \le 1$, then $\mathrm{M}(X,Y) := \mathrm{E}_{HD}(X,X) + \mathrm{E}_{HD}(Y,Y) - 2\mathrm{E}_{HD}(X,Y) \ge \sum_{i=1,j=1}^{n_1,n_2}(\sqrt{\mu^+_{i,j}} - \sqrt{\mu^-_{i,j}})^2 \ge 0$. Hence $\mathrm{M}(X,Y) = 0$ implies that

$$\mu^+_{i,j} = \mu^-_{i,j}, \ \sqrt{\mu^+_{i,j}}\sqrt{\mu^-_{i,j}}(\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j) - 1) = 0$$

for all i,j, then $\mu^+_{i,j}(\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j)-1) = 0$ for all i,j, which implies $\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j) = 1$ (i.e., $\mathbf{x}_i = \mathbf{y}_j$) for each pair (i,j) with $\mu^+_{i,j} > 0$. This statement in fact is equivalent to $X = Y$. In summary, $\mathrm{M}(X,Y) = 0$ implies $X = Y$.

B.2 Proof of the closest point property

Since $\|\mathbf{x}_i - \mathbf{y}_i\| \le \min(\|\mathbf{x}_i - \mathbf{y}_j\|, \|\mathbf{x}_j - \mathbf{y}_i\|)$, then

$$\exp(\frac{-\|\mathbf{x}_i - \mathbf{y}_i\|^2}{\sigma^2})\exp(\frac{-\|x_j - \mathbf{y}_j\|^2}{\sigma^2}) \ge \exp(\frac{-\|\mathbf{x}_i - \mathbf{y}_j\|^2}{\sigma^2})^2,$$

and then

$$\begin{aligned}
&E_{HD(\{\mu^+_{i,j},\mu^-_{i,j}\})}\\
&= \sum_{i=1,j=1}^{n,n}\sqrt{\mu^+_{i,j}\mu^-_{i,j}}\exp(-\|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2)\\
&\le \sum_{i=1,j=1}^{n,n}\sqrt{\mu^+_{i,j}\exp(-\frac{\|\mathbf{x}_i - \mathbf{y}_i\|^2}{\sigma^2})}\\
&\quad\sqrt{\mu^-_{i,j}\exp(-\frac{\|x_j - \mathbf{y}_j\|^2}{\sigma^2})},\\
&\le (\sum_{i=1,j=1}^{n,n}\mu^+_{i,j}\exp(-\frac{\|\mathbf{x}_i - \mathbf{y}_i\|^2}{\sigma^2}))^{1/2}\\
&\quad(\sum_{i=1,j=1}^{n,n}\mu^-_{i,j}\exp(-\frac{\|x_j - \mathbf{y}_j\|^2}{\sigma^2})^{1/2},\\
&= (\sum_{i=1}^{n}\exp(-\|\mathbf{x}_i - \mathbf{y}_i\|^2/\sigma^2))^{1/2}\\
&\quad(\sum_{j=1}^{n}\exp(-\|x_j - \mathbf{y}_j\|^2/\sigma^2)^{1/2}\\
&= \sum_{i=1,j=1}^{n,n}\sqrt{\delta_{i,j}\delta_{i,j}}\exp(-\|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2)\\
&= \mathrm{E}_{HD}(\{\delta_{i,j},\delta_{i,j}\}).
\end{aligned}$$

The second inequality comes from Cauchy-Schwartz inequalities. Thus, $\{\mu^+_{i,j} = \delta_{i,j} = \mu^-_{i,j} : i,j\}$ is a maximizer. Done.
One can also prove this by the duality. Let $a_i = \exp(-\|\mathbf{x}_i - \mathbf{y}_i\|^2/\sigma^2) = b_i$ for all i, then one can easily check that the selection of $(a_i, b_i, \mu^+_{i,j}, \mu^-_{i,j})$ satisfies conditions $(i), (ii), (iii)$ in subsection A.3.

B.3 Proof of the asymptotic case

More precisely, we prove the following statement: Let

$$(\mathbf{M}^+_\sigma, \mathbf{M}^-_\sigma) := arg\max_{\mathbf{M}^+,\mathbf{M}^-}\sum_{i=1,j=1}^{n,n}\sqrt{\mu^+_{i,j}\mu^-_{i,j}}\exp(-\|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2).$$

For a sequence $\{\sigma_k : k \in \mathbb{N}\}$ approaching infinity, we consider a converging sequence from $\{(\mathbf{M}^+_{\sigma_k}, \mathbf{M}^-_{\sigma_k})\}$(by taking a subsequence). Let $(\mathbf{M}^+, \mathbf{M}^-)$ be the limit. Then $\mathbf{M}^+ = \mathbf{M}^-$ and $\mathbf{M}^+$ is a minimizer of the Monge-Kantorovich mass transport problem[16]: $\min_{\mathbf{M}^+}\sum_{i=1,j=1}^{n,n}\mu^+_{i,j}\|\mathbf{x}_i - \mathbf{y}_j\|^2$.

Proof Substituting the expansion $\exp(-\|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2) = 1 - \|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2 + O(1/\sigma^4)$ as $\sigma \to \infty$ into E_{HD}, the original maximization problem becomes

$$\max_{\mathbf{M}^+,\mathbf{M}^-}\sum_{i=1,j=1}^{n,n}\sqrt{\mu^+_{i,j}\mu^-_{i,j}}(1 - \|\mathbf{x}_i - \mathbf{y}_j\|^2/\sigma^2) + O(1/\sigma^4),$$

which is equivalent to two optimization problems:

$$\max_{\mathbf{M}^+,\mathbf{M}^-} \sum_{i=1,j=1}^{n,n} \sqrt{\mu^+_{i,j}\mu^-_{i,j}} \tag{29}$$

and

$$\min_{\mathbf{M}^+,\mathbf{M}^-} \sum_{i=1,j=1}^{n,n} \sqrt{\mu^+_{i,j}\mu^-_{i,j}}\|\mathbf{x}_i - \mathbf{y}_j\|^2, \tag{30}$$

because $\sum_{i=1,j=1}^{n,n} \sqrt{\mu^+_{i,j}\mu^-_{i,j}}\|\mathbf{x}_i - \mathbf{y}_j\|^2$, and $\sum_{i=1,j=1}^{n,n} \sqrt{\mu^+_{i,j}\mu^-_{i,j}}\|\mathbf{x}_i - \mathbf{y}_j\|^4$ are bounded. In other words, the limit $(\mathbf{M}^+, \mathbf{M}^-)$ satisfies Eq. (29,30). By Cauchy-Schwartz inequalities, the maximization problem (29) yields $\mu^+_{i,j} = c\mu^-_{i,j}$ for all i,j with a constant c. Because of the same cardinality of points, summing over i,j, we have $c = 1$, i.e., $\mathbf{M}^+ = \mathbf{M}^-$. Finally, the minimization problem (30) yields the Monge-Kantorovich mass transport problem. □

B.4 Matching pairs with one outlier

Let $X = \{\mathbf{x}_i : i = 1, \ldots, n\} \cup \{\mathbf{x}_{n+1} = \mathbf{x}\}$ with unit weight $\mu^+_1 = \ldots = \mu^+_n = 1$ and $\alpha > 0$. Let $Y = \{\mathbf{y}_i = \mathbf{x}_i + t : i = 1, \ldots, n\}$ with unit weight $\mu^-_j = 1$ for $j = 1, \ldots, n$. Suppose that $\mathbf{y}_i$ is the closed point of $\mathbf{x}_i$ in Y, and $\mathbf{x}_i$ is the closed point of $\mathbf{y}_i$ in X for $i = 1, \ldots, n$. Assume that $\mathbf{y}_n$ is the closed point of $\mathbf{x}$ in Y.

Choose

$$a_i = \begin{cases} \mathrm{K}(\mathbf{x}_i, \mathbf{y}_i), & \text{if } i < n; \\ \frac{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2}{\sqrt{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2 + \alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}}, & \text{if } i = n; \\ \frac{\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}{\sqrt{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2 + \alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}}, & \text{if } i = n+1, \end{cases}$$

$$b_j = \begin{cases} \mathrm{K}(\mathbf{x}_j, \mathbf{y}_j), & \text{if } j < n; \\ \sqrt{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2 + \alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}, & \text{if } j = n; \end{cases}$$

$$(\mu^+_{i,j}, \mu^-_{i,j}) = \begin{cases} (1, \frac{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2}{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2 + \alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}), & \text{if } (i,j) = (n,n); \\ (\alpha, \frac{a\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2 + \alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}), & \text{if } (i,j) = (n+1,n); \\ (\delta_{i,j}, \delta_{i,j}), & \text{otherwise}; \end{cases}$$

We shall show that if α satisfies

$$\sqrt{1 + \frac{\alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2}} \le \min\left(\frac{\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}{\mathrm{K}(\mathbf{x},\mathbf{y}_j)^2}, \frac{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)\mathrm{K}(\mathbf{x}_j,\mathbf{y}_j)}{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_j)^2}\right), \tag{31}$$

this choice $\{a_i, b_j, \mu^+_{i,j}, \mu^-_{i,j} : i,j\}$ is an optimal solution.

Proof To verify it, we need to check (i)$a_i\mu^+_{i,j} = b_j\mu^-_{i,j}$, which is straightforward; (ii) $\mathrm{K}(\mathbf{x}_i,\mathbf{y}_j)^2 \le a_i b_j$, which is also quite easy, except the following nontrivial cases:
Case(a) $i = n$, $j < n$. According to the closest point assumption, $\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)\mathrm{K}(\mathbf{x}_j,\mathbf{y}_j) > \mathrm{K}(\mathbf{x}_n,\mathbf{y}_j)^2$. Then by Eq. (31), we have

$$a_n b_j = \frac{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)\mathrm{K}(\mathbf{x}_j,\mathbf{y}_j)}{\sqrt{1 + \alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2/\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2}} \ge \mathrm{K}(\mathbf{x}_n,\mathbf{y}_j)^2;$$

Case(b) $i = n+1$, $j < n$. Note that $\mathrm{K}(\mathbf{x}_j,\mathbf{y}_j) = \mathrm{K}(\mathbf{x}_n,\mathbf{y}_n) = \exp(-t^2/\sigma^2)$, and the closest point assumption $\mathrm{K}(\mathbf{x},\mathbf{y}_n) > \mathrm{K}(\mathbf{x},\mathbf{y}_j)$, then by Eq. (31)

$$a_{n+1} b_j = \frac{\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2\mathrm{K}(\mathbf{x}_j,\mathbf{y}_j)}{\sqrt{\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2 + \alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}}$$

$$= \frac{\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2}{\sqrt{1 + \alpha\mathrm{K}(\mathbf{x},\mathbf{y}_n)^2/\mathrm{K}(\mathbf{x}_n,\mathbf{y}_n)^2}} \ge \mathrm{K}(\mathbf{x},\mathbf{y}_j)^2;$$

(iii) the feasibility of $\{\mu^+_{i,j}, \mu^-_{i,j} : i,j\}$ is straightforward.

In summary, if α satisfies the condition (31), the above choice for $\{a_i, b_j, \mu^+_{i,j}, \mu^-_{i,j} : i,j\}$ is a solution. □

Acknowledgements The author appreciates that David Wilson and Edward A. Geiser provided these cardiac contours, while the author were in Mathematics department, University of Florida.

References

1. D. P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, Mass, 1999.
2. P. Besl and N. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14:239–256, 1992.
3. F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 11(6):567–585, 1989.
4. P. Chen, Y. Chen, and M. Rao. Kullback Leibler divergence based curve matching method. *SSVM*, pages 813–824, 2007.
5. H. Chui and A. Rangarajan. A feature registration framework using mixture models. *IEEE workshop on MMBIA*, pages 190–197, 2000.
6. H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. *CVPR*, 2:44–51, 2000.
7. H. Chui and A. Rangarajan. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 89(2-3):114–141, 2003.
8. I. L. Dryden and K. V. Mardia. *Statistical shape analysis.* Wiley, New York, 1998.
9. L. C. Evans and W. Gangbo. Differential equations methods for the Monge-Kantorovich mass transfer problem. *Memo. AMS*, 137(653), 1999.
10. J. Glaunes, A. Trouve, and L. Younes. Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. *CVPR*, 2:712–718, 2004.
11. S. Granger and X. Pennec. Multi-scale EM-ICP: A fast and robust approach for surface registration. *ECCV*, 4:418–432, 2002.
12. R. A. Horn and C. R. Johnson. *Matrix theory.* Cambridge University Press, New York, NY, 1985.
13. P. Huber. *Robust Statistics.* John Wiley& Sons, New York,NY, 1981.
14. B. Jian and B. C. Vemuri. A robust algorithm for point set registration using mixture of Gaussians. *ICCV*, 2:1246–1251, 2005.
15. S. C. Joshi and M. I. Miller. Landmark matching via large deformation diffeomorphism. *IEEE Image Proc.*, 9(8):1357–1370, 2000.
16. L. V. Kantorovich. On a problem of Monge. *Uspekhi Mat. Nauk.*, 3:225–226, 1948.
17. M. Kass, A. Witkin, and D. Terzopoulos. Snake: active contour models. *International Journal of Computer Vision*, 1:321–331, 1988.
18. S. Kay. *Fundamentals of statistical signal processing.* Prentice Hall PTR, New Jersey, 1993.
19. P. A. Knight. The Sinkhorn–Knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.
20. M. Leordeanu and M. Hebert. A spectral technique for correspondence problems using pairwise constraints. *ICCV*, pages 1482–1489, 2005.
21. B. Luo and E. Hancock. A unified framework for alignment and correspondence. *Computer Vision and Image Understanding*, 92:26–55, 2003.
22. G. McNeil and S. Vijayakumar. A probabilistic approach to robust shape matching. *IEEE ICIP*, pages 937–940, 2006.
23. M. Miller and L. Younes. Group actions, homeomorphism, matching: A general framework. *International Journal of Computer Vision*, 41:61–84, 2001.
24. C. Papadimitriou and K. Steiglitz. *Combinatorial Optimization.* Prentice Hall PTR, New Jersey, 1982.

25. G. Scott and H. C. Longuet-Higgins. An algorithm for associating the features of two images. *Proceedings of the Royal Society London: Biological Sciences*, 244(1309):21–26, 1991.
26. T. Sebastian, P. Klein, and B. Kimia. On aligning curves. *IEEE Trans. Pattern Anal. Mach. Intell.*, 5:116–125, 2003.
27. I. Sethi and R. Jain. Finding trajectories of feature points in a monocular image sequence. *IEEE Trans. Pattern Anal. Mach. Intell.*, 9:56–73, 1987.
28. R. Sinkhorn. A relationship between arbitrary positive matrices and doubly stochastic matrices. *Ann. Math. Statist.*, 35:876–879, 1964.
29. D. W. Thompson. *ON GROWTH AND FORM*. Cambridge University Press, Cambridge, 1917.
30. L. N. Trefethen and D. Bau. *Numerical linear algebra.* SIAM, Philadelphia, PA, 1997.
31. Y. Tsin and T. Kanade. A correlation-based approach to robust point set registration. *ECCV*, 3:558–569, 2004.
32. Z. Tu, S. Zheng, and A. Yuille. Shape matching and registration by data-driven EM. *Computer Vision and Image Understanding*, 109(3):290 – 304, 2008.
33. S. Ullman. *The interpretation of visual motion*. MIT Press, Cambridge, MA, 1979.
34. G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, PA, 1990.
35. F. Wang, B. Vemuri, A. Rangarajan, and S. Eisenschenk. Simultaneous nonrigid registration of multiple point sets and atlas construction. *IEEE PAMI*, 30(11):2011–2022, 2008.
36. W. M. Wells. Statistical approaches to feature-based object recognition. *International Journal of Computer Vision*, 22:63–98, 1997.
37. L. Younes. Computable elastic distances between shapes. *SIAM on Appl. Math.*, 58(2):565–586, 1998.
38. Z. Zhang. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994.
39. C. L. Zitnick and T. Kanade. A cooperative algorithm for stereo matching and occlusion detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22:675–684, 2000.