Machine Learning with Applications 6 (2021) 100088

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Quantile–Quantile Embedding for distribution transformation and manifold embedding with ability to choose the embedding distribution

Benyamin Ghojogh [a,*], Fakhri Karray [b], Mark Crowley [a]

[a] Machine Learning Laboratory, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada
[b] Centre for Pattern Analysis and Machine Intelligence, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

ARTICLE INFO

Keywords:
Quantile–Quantile Embedding (QQE)
Quantile–quantile plot
Distribution transformation
Manifold embedding
Embedding distribution
Class discrimination

ABSTRACT

We propose a new embedding method, named Quantile–Quantile Embedding (QQE), for distribution transformation and manifold embedding with the ability to choose the embedding distribution. QQE, which uses the concept of quantile–quantile plot from visual statistical tests, can transform the distribution of data to any theoretical desired distribution or empirical reference sample. Moreover, QQE gives the user a choice of embedding distribution in embedding the manifold of data into the low dimensional embedding space. It can also be used for modifying the embedding distribution of other dimensionality reduction methods, such as PCA, t-SNE, and deep metric learning, for better representation or visualization of data. We propose QQE in both unsupervised and supervised forms. QQE can also transform a distribution to either an exact reference distribution or its shape. We show that QQE allows for better discrimination of classes in some cases. Our experiments on different synthetic and image datasets show the effectiveness of the proposed embedding method.

1. Introduction

Regardless of the data science or machine learning task, there is no doubt that the distribution of the available data instances is highly relevant to the final outcome of any algorithm. This distribution may be a standard distribution, such as a Gaussian, or it may be some other more exotic distribution or it may be unknown to us entirely. Now, while pre-processing of data to remove noise, normalize, or otherwise adjust the data to be appropriate for a given algorithm is very common, imagine that we know, or suspect, that the distribution of the given data may not be suitable for the target purpose. The reason for this could be anything, including knowledge about bias in the datasets, an incompatible distribution type for maximizing class discrimination or for representation and interpretation reasons. In these cases, we suggest it would be useful to be able to transform the distribution of dataset, as a pre-processing step, to a known distribution form while maintaining the original local relationships and distances between data instances so that the unique character of the dataset is maintained (Saul & Roweis, 2003). What we present here is a general approach for doing just that, and we include a number of variants and experimental demonstrations of the utility and effectiveness of the approach.

For this distribution transformation, one could try to make all moments of data equal to the moments of the desired distribution (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2007, 2012) but this quickly becomes computationally expensive. Furthermore, moments of non-standard distributions can be hard to compute in some cases. Another problem with simply matching all moments is that it results in transformation to the *exact* desired distribution rather than the "shape" of the desired distribution which is more desirable. Note that transformation of data to the shape of another distribution means that the general shape of the Probability Density Function (PDF) of data becomes similar to the desired PDF regardless of the mean and scale of the distribution. One could also imagine, that the desired distribution may only be known indirectly via the distribution of some other empirical reference sample. Thus, our method for distribution transformation should support providing a desired distribution either as a theoretical PDF/Cumulative Distribution Function (CDF) or as an empirical reference sample.

A further connection for our proposed approach would be allowing a choice of distribution of embedded data in the field of manifold learning and dimensionality reduction. That is, we might want to choose what distribution the data instances will have after being embedded by a dimensionality reduction method. In dimensionality reduction, the choice of embedding distribution is usually not given to the user but it is very relevant as some dimensionality reduction methods already make assumptions about the distribution of neighbors of data points.

The code (and data) in this article has been certified as Reproducible by Code Ocean: (https://codeocean.com/). More information on the Reproducibility Badge Initiative is available at https://www.elsevier.com/physical-sciences-and-engineering/computer-science/journals.

* Corresponding author.
E-mail addresses: bghojogh@uwaterloo.ca (B. Ghojogh), karray@uwaterloo.ca (F. Karray), mcrowley@uwaterloo.ca (M. Crowley).

https://doi.org/10.1016/j.mlwa.2021.100088
Received 20 August 2020; Received in revised form 18 June 2021; Accepted 21 June 2021
Available online 29 June 2021
2666-8270/© 2021 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Meanwhile, other methods do not even make any assumption on the embedding distribution and yet do not give any choice of embedding distribution to the user. We will enumerate some examples for these methods in Section 2.

In this paper, we propose a new embedding method, named Quantile–Quantile Embedding (QQE), which can be used for distribution transformation and manifold learning with a user-specified choice of embedding distribution. The features and advantages of QQE are summarized as follows:

1. Distribution transformation to a desired distribution either as a PDF/CDF or an empirical reference distribution given by user. The entire dataset can be transformed in an unsupervised manner or every class in the dataset can be transformed in a supervised manner.
2. Manifold embedding of high dimensional data into a lower dimensional embedding space with the choice of embedding distribution by the user. Manifold embedding in QQE can also modify the embedding of other manifold learning methods, such as Principal Component Analysis (PCA), Fisher Discriminant Analysis (FDA), Student-t distributed Stochastic Neighbor Embedding (t-SNE), Locally Linear Embedding (LLE), and deep metric learning, for better discrimination of classes or better representation/visualization of data.
3. For both distribution transformation and manifold embedding tasks, the distribution can be transformed to either the exact desired distribution or merely the shape of it. One of the many applications of exact distribution transformation is separation of classes in data.

The remainder of this paper is organized as follows. We review the related work in Section 2. Section 3 introduces the technical background on quantile functions, the univariate quantile–quantile plot, and its multivariate version. In Section 4, we propose the QQE method for both distribution transformation and manifold embedding. The experimental results are reported in Section 5. Finally, Section 6 concludes the paper and enumerates the future directions.

2. Related work

2.1. Methods for difference of distributions

An obvious connection exists between this work and the task of computing the difference between two distributions which is a rich field in statistics. One of the most well-known methods is the Kullback–Leibler (KL) divergence (Kullback & Leibler, 1951). KL-divergence, which is a relative entropy from one distribution to the other one, has been widely used in deep learning (Goodfellow, Bengio, & Courville, 2016). Another measure for difference of distributions of two random variables is Maximum Mean Discrepancy (MMD) or kernel two-sample test. It is a measure of difference of two distributions by comparing their moments (Gretton et al., 2007, 2012). This comparison of moments can be performed after pulling data to the feature space using kernels (Hofmann, Schölkopf, & Smola, 2008). MMD uses distances in the feature space (Schölkopf, 2001). It has been used in machine learning algorithms such as generative moment matching networks (Li, Swersky, & Zemel, 2015; Ren, Zhu, Li, & Luo, 2016). Another measure for measuring the relation of two random variables is Hilbert–Schmidt Independence Criterion (HSIC) (Gretton, Bousquet, Smola, & Schölkopf, 2005). Calculating the dependence of two random variables is difficult while calculating the linear dependence, named correlation, is much simpler. Therefore, for computation of dependence of two random variables, HSIC pulls data to the feature space using kernels (Hofmann et al., 2008) and then computes the correlation between them in that space. This correlation is a good estimate for the dependence in the input space. Two example uses of HSIC in machine learning are supervised PCA (Barshan, Ghodsi, Azimifar, & Jahromi, 2011) and supervised guided LLE (Alipanahi & Ghodsi, 2011). Note that the formulas of the three introduced methods for measuring the difference of distributions will be provided in Section 5.1. We have used these measures for quantitatively discussing the results of QQE algorithm.

2.2. Quantile plots for visual statistical tests

The quantile function for a distribution is defined as the inverse of the CDF (Hyndman & Fan, 1996; Parzen, 1979). If we plot the quantile function, we will have the quantile plot (Galton et al., 1885). There are multivariate versions of quantile plots (Chaudhuri, 1996) where data instances are multivariate rather than univariate. In case there are two sets of data instances with two distributions, one can match the quantile plots of these two datasets and have the quantile–quantile plot or qq-plot (Loy, Follett, & Hofmann, 2016). Using the qq-plot, statisticians can visually test whether the two distributions are equal and if not, how different they are (Loy et al., 2016; Oldford, 2016). There also exist multivariate versions of qq-plot such as fuzzy qq-plot (Easton & McCulloch, 1990). These multivariate qq-plots can be used for visual assessment of whether two distributions match or not. The technical required background on quantile plot and qq-plot are provided in Section 3.

2.3. Embedding distribution in manifold learning methods

Some manifold learning and dimensionality reduction methods make an assumption about the distribution of neighbors of data points. For example, Stochastic Neighbor Embedding (SNE) and t-SNE take the Gaussian distribution (Hinton & Roweis, 2003) and Cauchy (Maaten & Hinton, 2008) (or Student-t (Van Der Maaten, 2009)) distribution for the neighborhood of points, respectively. These methods make some strong assumptions about the neighborhood of points and do not give freedom of choice to the user for the embedding distribution. Some manifold learning methods, however, do not even make any assumption about the embedding distribution and yet do not give any choice of embedding distribution to the user. Some examples are PCA (Ghojogh & Crowley, 2019), Multi-dimensional Scaling (MDS) (Cox & Cox, 2008), Sammon mapping (Sammon, 1969), FDA (Ghojogh, Karray, & Crowley, 2019), Isomap (Tenenbaum, De Silva, & Langford, 2000), LLE (Roweis & Saul, 2000; Saul & Roweis, 2003), and deep manifold learning (He, Zhang, Ren, & Sun, 2016; Schroff, Kalenichenko, & Philbin, 2015). Note that some of these methods make assumptions but not as a distribution for the embedding. For example, FDA assumes the Gaussian distribution for data in the input space and LLE assumes just unit covariance and zero mean for the embedded data.

3. Quantile and quantile–quantile plots

3.1. Quantile function and quantile plot

The *quantile function* for a distribution is defined as (Hyndman & Fan, 1996; Parzen, 1979):

$$Q(p) := F^{-1}(p) := \inf\{x \mid F(x) \geq p\}, \tag{1}$$

where $p \in [0, 1]$ is called *position* and $F(x)$ is the CDF. The quantile function can also be defined as:

$$Q(p) := \arg\min_{\theta \in \mathbb{R}} \mathbb{E}\big[|X - \theta| + (2p - 1)(X - \theta)\big], \tag{2}$$

where X is a random variable with $\mathbb{E}[X] < \infty$ (Ferguson, 1967; Serfling, 2004). The two-dimensional plot $(p, Q(p))$ is called the *quantile plot* which was first proposed by Sir Francis Galton (Galton et al., 1885). Its name was *ogival curve* primarily as it was like an ogive because of the normal distribution of his measured experimental sample.

If we have a drawn sample, with sample size n from a distribution, the quantile plot is a *sample (or empirical) quantile*. The sample quantile

plot is $(p_i, Q(p_i)), \forall i \in \{1, \dots, n\}$. For the sample quantile, we can determine the ith position, denoted by p_i, as:

$$p_i := \frac{i - \alpha}{n - \alpha - \beta + 1}, \tag{3}$$

where different values for α and β result in different positions (Leon Harter, 1984). The simplest type of position is $p_i = i/n$ (with $\alpha = \beta = 0$) (Parzen, 1979). The most well-known position is $p_i = (i-0.5)/n$ (with $\alpha = 0.5, \beta = 0$) (Allen, 1914). However, it is suggested in Hyndman and Fan (1996) to use $p_i = (i - 1/3)/(n + 1/3)$ (with $\alpha = \beta = 1/3$) which is median unbiased (Reiss, 2012). It is noteworthy that Galton also suggested that we can measure the quantile function only in $p \in \{0.02, 0.09, 0.25, 0.50, 0.75, 0.91, 0.98\}$ as a summary (Galton, 1874). His summary is promising only for the normal distribution; however, with the power of today's computers we can compute the sample quantile with fine steps.

For the multivariate quantile plot, *spatial rank* fulfills the role played by position in the univariate case. Spatial rank $\boldsymbol{u}_i \in \mathbb{R}^d$ of $\boldsymbol{x}_i \in \mathbb{R}^d$ with respect to the sample $\{\boldsymbol{x}_j\}_{j=1}^n$ is defined as (Dhar, Chakraborty, & Chaudhuri, 2014; Marden, 2004; Möttönen & Oja, 1995; Serfling, 2004):

$$\boldsymbol{u}_i := \frac{1}{n} \sum_{j=1, j\neq i}^{n} \frac{\boldsymbol{x}_i - \boldsymbol{x}_j}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2}, \tag{4}$$

whose term in the summation is a generalization of the sign function for the multivariate vector (Marden, 2004). Eq. (2) can be restated as $\arg\min_\theta \mathbb{E}(|X - \theta| + u(X - \theta))$ where $[-1, 1] \ni u := 2p - 1$ (Chaudhuri, 1996). The multivariate *spatial quantile* (or *geometrical quantile*) for the multivariate spatial rank $\boldsymbol{u} \in \mathbb{R}^d$ is defined as:

$$\boldsymbol{Q}(\boldsymbol{u}) := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathbb{E}(\boldsymbol{\Phi}(\boldsymbol{u}, \boldsymbol{x} - \boldsymbol{\theta}) - \boldsymbol{\Phi}(\boldsymbol{u}, \boldsymbol{x})), \tag{5}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is a random vector, $\boldsymbol{\Phi}(\boldsymbol{u}, \boldsymbol{t}) := \|\boldsymbol{t}\|_2 + \boldsymbol{u}^\top \boldsymbol{t}$, and $\boldsymbol{u}$ is a vector in unit ball, i.e., $\boldsymbol{u} \in \{\boldsymbol{v} \mid \boldsymbol{v} \in \mathbb{R}^d, \|\boldsymbol{v}\|_2 < 1\}$ (Chaudhuri, 1996; Dhar et al., 2014; Serfling, 2004).

3.2. Quantile–quantile plot

Assume we have two quantile functions for two univariate distributions. If we match their positions and plot $(Q_1(p), Q_2(p)), \forall p \in [0, 1]$, we will have *quantile–quantile plot* or *qq-plot* in short (Loy et al., 2016). Again, this plot can be an empirical plot, i.e., $(Q_1(p_i), Q_2(p_i)), \forall i \in \{1, \dots, n\}$. Note that the qq-plot is equivalent to the quantile plot for the uniform distribution as we have $Q(p) = p$ in this distribution. Usually, as a statistical test, we want to see whether the first distribution is similar to the second empirical or theoretical distribution (Loy et al., 2016); therefore, we refer to the first and second distributions as the *observed and reference distributions*, respectively (Easton & McCulloch, 1990). Note that if the qq-plot of two distributions is a line with slope 1 (angle $\pi/4$) and intercept 0, the two distributions have the same distributions (Oldford, 2016). The slope and the intercept of the line show the difference of spread and location of the two distributions (Loy et al., 2016).

In order to extend the qq-plot to multivariate distributions, we can consider the marginal quantiles. However, this fails to take the dependence of marginals into account (Dhar et al., 2014; Easton & McCulloch, 1990). There are several existing methods for a promising generalization. One of these methods is *fuzzy qq-plot* (Easton & McCulloch, 1990) (note that it is not related to fuzzy logic). In a fuzzy qq-plot, a sample of size n is drawn from the reference distribution and the data points of the two samples are matched using optimization. An affine transformation is also applied to the observed sample in order to have an invariant comparison to the affine transformation. In the multivariate qq-plot, the matched data points are used to plot the qq-plots for every component; therefore, we will have d qq-plots where d is the dimensionality of data. Note that these plots are different from the d qq-plots for the marginal distributions. The technical details of fuzzy qq-plot is explained in the following.

3.3. Multivariate fuzzy quantile–quantile plot

Assume we have a dataset with size n and dimensionality d, i.e., $\{\boldsymbol{x}_i \in \mathbb{R}^d\}_{i=1}^n$. We want to transform its distribution as $\boldsymbol{x}_i \mapsto \boldsymbol{y}_i, \forall i \in \{1, \dots, n\}$. We draw a sample $\{\boldsymbol{y}_i \in \mathbb{R}^d\}_{i=1}^n$ of size n from the desired (reference) distribution. Note that in case we already have a reference sample $\{\boldsymbol{y}_i \in \mathbb{R}^d\}_{i=1}^m$ rather than the reference distribution, we can employ bootstrapping or oversampling if $m > n$ and $m < n$, respectively, to have $m = n$. We match the data points $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$ (Easton & McCulloch, 1990):

$$\underset{\boldsymbol{A}, \boldsymbol{b}, \sigma}{\text{minimize}} \quad \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{y}_{\sigma(i)} - \boldsymbol{b}\|_2^2, \tag{6}$$

where $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{b} \in \mathbb{R}^d$ are used to make the matching problem invariant to affine transformation. If $\mathcal{P}$ is the set of all possible permutations of integers $\{1, \dots, n\}$, we have $\sigma \in \mathcal{P}$. This optimization problem finds the best permutation regardless of any affine transformation. Note that one can exchange $\boldsymbol{x}$ and $\boldsymbol{y}$ in Eq. (6) and the following equations to match $\boldsymbol{x}$ with $\boldsymbol{y}$. As long as matching is performed correctly, that is fine.

In order to solve this problem, we iteratively switch between solving for $\boldsymbol{A}$, $\boldsymbol{b}$, and σ until there is no change in σ (Easton & McCulloch, 1990). Given $\boldsymbol{A}$ and $\boldsymbol{b}$, we solve:

$$\min_{\sigma} \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{y}_{\sigma(i)} - \boldsymbol{b}\|_2^2 \equiv \min_{\boldsymbol{\Psi}} \sum_{i=1}^{n} \sum_{j=1}^{n} \boldsymbol{C}(i, j) \boldsymbol{\Psi}(i, j), \tag{7}$$

which is an assignment problem and can be solved using the Hungarian method (Kuhn, 1955). $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$ are the cost matrix and a matrix with only one 1 in every row, respectively. Note that $\boldsymbol{\Psi}(i, j) = 1$ means that $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are matched. $\boldsymbol{C}$ should be computed before solving the optimization where $\boldsymbol{C}(i, j) := \|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{y}_j - \boldsymbol{b}\|_2^2$.

According to the 1's in the obtained $\boldsymbol{\Psi}$, we have σ. Then given σ, we solve:

$$\underset{\boldsymbol{A}, \boldsymbol{b}}{\text{minimize}} \quad \sum_{i=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{y}_{\sigma(i)} - \boldsymbol{b}\|_2^2, \tag{8}$$

which is a multivariate regression problem. The solution is (Hastie, Tibshirani, & Friedman, 2009):

$$\mathbb{R}^{(d+1) \times d} \ni \boldsymbol{\beta} := (\check{\boldsymbol{Y}}^\top \check{\boldsymbol{Y}})^{-1} \check{\boldsymbol{Y}}^\top \check{\boldsymbol{X}}, \tag{9}$$

where $\mathbb{R}^{n \times (d+1)} \ni \check{\boldsymbol{Y}} := \big[[\boldsymbol{y}_{\sigma(1)}, \dots, \boldsymbol{y}_{\sigma(n)}]^\top, \boldsymbol{1}_{n \times 1}\big]$ and $\mathbb{R}^{n \times d} \ni \check{\boldsymbol{X}} := \boldsymbol{X}^\top = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^\top$. We will have $\boldsymbol{\beta} = [\boldsymbol{A}, \boldsymbol{b}]^\top$. Therefore, $\boldsymbol{A}$ and $\boldsymbol{b}$ are found where $\boldsymbol{A}^\top$ is the top $d \times d$ sub-matrix of $\boldsymbol{\beta}$ and $\boldsymbol{b}^\top$ is the last row of $\boldsymbol{\beta}$.

Note that it is better to set the initial rotation matrix to the identity matrix, i.e. $\boldsymbol{A}^{(0)} = \boldsymbol{I}$, to reduce the amount of assignment rotation. In this way, only a few iterations will suffice to solve the matching problem. This iterative optimization gives us the matching σ and the samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$ are matched. Then, we have d qq-plots, one for every dimension. These qq-plots are named fuzzy qq-plots (Easton & McCulloch, 1990). Considering the spatial ranks, the quantiles are (Dhar et al., 2014):

$$\boldsymbol{Q}_X(\boldsymbol{u}_i) = \boldsymbol{x}_i, \qquad \forall i \in \{1, \dots, n\}, \tag{10}$$

$$\boldsymbol{Q}_Y(\boldsymbol{u}_i) = \boldsymbol{y}_{\sigma(i)}, \quad \forall i \in \{1, \dots, n\}. \tag{11}$$

4. Quantile–quantile embedding

In QQE, we want to transform data instances from one dataset $\{\boldsymbol{x}_i^0\}_{i=1}^n$ to another dataset $\{\boldsymbol{x}_i\}_{i=1}^n$ where distribution is transformed to a desired target distribution while the local relationships between points are preserved as much as possible. Formally, we define the task of distribution transformation as follows:

Definition 1 (*Distribution Transformation*)**.** For a sample $\{\boldsymbol{x}_i^0\}_{i=1}^n$ of size n in $\mathbb{R}^d$ space, the mapping $\boldsymbol{x}_i^0 \mapsto \boldsymbol{x}_i, \forall i \in \{1, \dots, n\}$ is a distribution transformation where the distribution of $\{\boldsymbol{x}_i\}_{i=1}^n$ is the known desired distribution and the local distances of nearby points in $\{\boldsymbol{x}_i^0\}_{i=1}^n$ are preserved in $\{\boldsymbol{x}_i\}_{i=1}^n$ as much as possible.

Distribution transformation can be performed in two ways: (i) the distribution of data is transformed to the "exact" reference distribution, and (ii) only the "shape" of the reference distribution is considered to transform to. In the following subsections, we detail these two approaches then we introduce a manifold embedding variation, and finally explain the use of unsupervised and supervised approaches for QQE.

4.1. Distribution transformation to exact reference distribution

QQE can be used for transformation of data to some exact reference distribution where all moments of the data become equal to the moments of the reference distribution. We start with an initial sample $\{\boldsymbol{x}_i^0\}_{i=1}^n$ and transform it to $\{\boldsymbol{x}_i\}_{i=1}^n$ whose distribution is desired to be the same as the distribution of a reference sample $\{\boldsymbol{y}_{\sigma(i)}\}_{i=1}^n$ or a reference distribution. For this, we consider the fuzzy qq-plot of $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_{\sigma(i)}\}_{i=1}^n$. When the d qq-plots are obtained by the fuzzy qq-plot, we can use them to embed the data for distribution transformation. Therefore, the qq-plot of every dimension should be a line with slope one and intercept zero (Oldford, 2016). Let $Q_l(\boldsymbol{u}_i) \in \mathbb{R}$ denote the lth dimension of $\mathbb{R}^d \ni \boldsymbol{Q}(\boldsymbol{u}_i) = [Q_1(\boldsymbol{u}_i), \dots, Q_d(\boldsymbol{u}_i)]^\top$ which is used for the ith data point in the lth qq-plot. Consider $Q_l(\boldsymbol{u}_i)$ for the matched data and the reference sample, denoted by $Q_{X,l}(\boldsymbol{u}_i)$ and $Q_{Y,l}(\boldsymbol{u}_i)$, respectively. In order to have the line in the qq-plot, we should minimize $\sum_{i=1}^n \sum_{l=1}^d \big(Q_{X,l}(\boldsymbol{u}_i) - Q_{Y,l}(\boldsymbol{u}_i)\big)^2$. According to Eqs. (10) and (11), this cost function is equivalent to $\sum_{i=1}^n \sum_{l=1}^d (x_{i,l} - y_{\sigma(i),l})^2$ where $x_{i,l}$ and $y_{\sigma(i),l}$ denote the lth dimension of $\boldsymbol{x}_i = [x_{i,1}, \dots, x_{i,d}]^\top$ and $\boldsymbol{y}_{\sigma(i)} = [y_{\sigma(i),1}, \dots, y_{\sigma(i),d}]^\top$, respectively. In vector form, the cost function is restated as:

$$\mathcal{L}_1 := \frac{1}{2} \sum_{i=1}^n \|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|_2^2. \tag{12}$$

On the other hand, according to our definition of distribution transformation, we should also preserve the local distances of the nearby data points as far as possible to embed data locally (Saul & Roweis, 2003). For preserving the local distances, we minimize the differences of local distances between data and transformed data. We use the k-nearest neighbors (k-NN) graph for the set $\{\boldsymbol{x}_i\}_{i=1}^n$. Let $\mathcal{N}_i$ denote the set containing the indices of the k neighbors of $\boldsymbol{x}_i$. The cost to be minimized is:

$$\mathcal{L}_2 := \frac{1}{2a} \sum_{i=1}^n \sum_{j \in \mathcal{N}_i} w_{ij} \big(d_x(i,j) - d_x^0(i,j)\big)^2, \tag{13}$$

where $d_x(i,j) := \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$, $d_x^0(i,j) := \|\boldsymbol{x}_i^0 - \boldsymbol{x}_j^0\|_2$, and $a := \sum_{i=1}^n \sum_{j \in \mathcal{N}_i} d_x^0(i,j)$ is the normalization factor. The weight $w_{ij} := 1/d_x^0(i,j)$ gives more value to closer points as expected. Note that if $k = n-1$, Eq. (13) is the cost function used in Sammon mapping (Lee & Verleysen, 2007; Sammon, 1969). We use this cost as a regularization term in our optimization. Therefore, our optimization problem is:

$$\underset{\boldsymbol{X}}{\text{minimize}} \quad \mathcal{L} := \frac{1}{2} \sum_{i=1}^n \Big(\|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|_2^2 + \frac{\lambda}{a} \sum_{j \in \mathcal{N}_i} w_{ij} \big(d_x(i,j) - d_x^0(i,j)\big)^2 \Big), \tag{14}$$

where $\lambda > 0$ is the regularization parameter.

Proposition 1. *The gradient of the cost function with respect to $x_{i,l}$ is:*

$$\frac{\partial \mathcal{L}}{\partial x_{i,l}} = (x_{i,l} - y_{\sigma(i),l}) + \frac{\lambda}{a} \sum_{j \in \mathcal{N}_i} \frac{d_x(i,j) - d_x^0(i,j)}{d_x(i,j)\, d_x^0(i,j)} (x_{i,l} - x_{j,l}). \tag{15}$$

Proof. Proof in Appendix A. □

Proposition 2. *The second derivative of the cost function with respect to $x_{i,l}$ is:*

$$\frac{\partial^2 \mathcal{L}}{\partial x_{i,l}^2} = 1 + \frac{\lambda}{a} \sum_{j \in \mathcal{N}_i} \Big(\frac{d_x(i,j) - d_x^0(i,j)}{d_x(i,j)\, d_x^0(i,j)} + \frac{(x_{i,l} - x_{j,l})^2}{\big(d_x(i,j)\big)^3} \Big). \tag{16}$$

Proof. Proof in Appendix B. □

We use the quasi-Newton's method (Nocedal & Wright, 2006) for solving this optimization problem inspired by (Sammon, 1969). If we consider the vectors component-wise, the diagonal quasi-Newton's method updates the solution as (Lee & Verleysen, 2007):

$$x_{i,l}^{(v+1)} := x_{i,l}^{(v)} - \eta \Big| \frac{\partial^2 \mathcal{L}}{\partial x_{i,l}^2} \Big|^{-1} \frac{\partial \mathcal{L}}{\partial x_{i,l}}, \tag{17}$$

$\forall i \in \{1, \dots, n\}, \forall l \in \{1, \dots, d\}$, where v is the index of iteration, $\eta > 0$ is the learning rate, and $|.|$ denotes the absolute value guaranteeing that we move toward the minimum and not maximum in the Newton's method.

4.2. Distribution transformation to the shape of reference distribution

In distribution transformation, we can ignore the location and scale of the reference distribution and merely change the distribution of the observed sample to look like the "shape" of the reference distribution regardless of its location and scale. In other words, we start with an initial sample $\{\boldsymbol{x}_i^0\}_{i=1}^n$ and transform it to $\{\boldsymbol{x}_i\}_{i=1}^n$ whose shape of distribution is desired to be similar to the shape of distribution of a reference sample $\{\boldsymbol{y}_{\sigma(i)}\}_{i=1}^n$. Recall that if the qq-plot is a line, the shapes of the distributions are the same where the intercept and slope of the line correspond to the location and scale (Oldford, 2016). Therefore, in our optimization, rather than trying to make the qq-plot a line with slope one and intercept zero, we try to make it the closest line possible with any slope and intercept. This line can be found by fitting a line as a least squares problem, i.e., a linear regression problem. For the qq-plot of every dimension, we fit a line to the qq-plot. If we define $\mathbb{R}^n \ni \check{\boldsymbol{Q}}_{Y,l} := [Q_{Y,l}(\boldsymbol{u}_1), \dots, Q_{Y,l}(\boldsymbol{u}_n)]^\top$, let $\mathbb{R}^{n \times 2} \ni \boldsymbol{\Gamma}_l := [\boldsymbol{1}_{n \times 1}, \check{\boldsymbol{Q}}_{Y,l}]$. Fitting a line to the qq-plot of the lth dimension is the following least squares problem:

$$\underset{\boldsymbol{\beta}_l}{\text{minimize}} \quad \frac{1}{2} \Big\| \boldsymbol{Q}_X(\boldsymbol{u}_i) - \boldsymbol{\Gamma}_l\, \boldsymbol{\beta}_l \Big\|_2^2 \overset{(10)}{=} \frac{1}{2} \Big\| \boldsymbol{x}_l - \boldsymbol{\Gamma}_l\, \boldsymbol{\beta}_l \Big\|_2^2, \tag{18}$$

whose solution is (Hastie et al., 2009):

$$\mathbb{R}^2 \ni \boldsymbol{\beta}_l = (\boldsymbol{\Gamma}_l^\top \boldsymbol{\Gamma}_l)^{-1} \boldsymbol{\Gamma}_l^\top \boldsymbol{x}_l, \tag{19}$$

where $\mathbb{R}^n \ni \boldsymbol{x}_l := [x_{1,l}, \dots, x_{n,l}]^\top$. The n points on the line fitted to the qq-plot of the lth dimension are:

$$\mathbb{R}^n \ni \boldsymbol{\mu}_l := \boldsymbol{\Gamma}_l\, \boldsymbol{\beta}_l = [\mu_{\sigma(1),l}, \dots, \mu_{\sigma(n),l}]^\top, \tag{20}$$

which are used instead of $\boldsymbol{Q}_Y(\boldsymbol{u}_i), \forall i$ in our optimization. Defining $\mathbb{R}^d \ni \check{\boldsymbol{\mu}}(\boldsymbol{y}_{\sigma(i)}) := [\mu_{\sigma(i),1}, \dots, \mu_{\sigma(i),d}]^\top$, the optimization problem is:

$$\underset{\boldsymbol{Y}}{\text{minimize}} \quad \mathcal{L} := \frac{1}{2} \sum_{i=1}^n \Big(\|\boldsymbol{x}_i - \check{\boldsymbol{\mu}}(\boldsymbol{y}_{\sigma(i)})\|_2^2 + \frac{\lambda}{a} \sum_{j \in \mathcal{N}_i} w_{ij} \big(d_x(i,j) - d_x^0(i,j)\big)^2 \Big). \tag{21}$$

Similar to Proposition 1, the gradient is:

$$\frac{\partial \mathcal{L}}{\partial x_{i,l}} = (x_{i,l} - \mu_{\sigma(i),l}) + \frac{\lambda}{a} \sum_{j \in \mathcal{N}_i} \frac{d_x(i,j) - d_x^{(0)}(i,j)}{d_x(i,j)\, d_x^{(0)}(i,j)} (x_{i,l} - x_{j,l}), \tag{22}$$

and the second derivative is the same as Proposition 2. We again solve the optimization using the diagonal quasi-Newton's method (Nocedal & Wright, 2006).

Table 1
The runtime for experiments in this paper. The reported times sum the timings for matching and fuzzy qq-plot iterations. In QQE manifold embedding, the time for initialization is not included. All times are in seconds. Letters U and S denote unsupervised and supervised QQE approaches, respectively.

Experiment	Fig. 1 (1st row)	Fig. 1 (2nd row)	Fig. 1 (3rd row)	Fig. 1 (4th row)	Fig. 2	Fig. 4		
Time	206.56	253.49	364.08	278.76	51.57	1064.85		
Experiment	Fig. 5 (PCA, U)	Fig. 5 (PCA, S)	Fig. 5 (FDA, U)	Fig. 5 (FDA, S)	Fig. 5 (Isomap, U)	Fig. 5 (Isomap, S)	Fig. 5 (LLE, U)	Fig. 5 (LLE, S)
Time	206.81	450.87	262.82	269.86	187.53	283.65	328.74	276.10
Experiment	Fig. 5 (t-SNE, U)	Fig. 5 (t-SNE, S)	Fig. 5 (S, Exact)					
Time	289.68	289.62	365.24					
Experiment	Fig. 6 (PCA, U)	Fig. 6 (PCA, S)	Fig. 6 (FDA, U)	Fig. 6 (FDA, S)	Fig. 6 (Isomap, U)	Fig. 6 (Isomap, S)	Fig. 6 (LLE, U)	Fig. 6 (LLE, S)
Time	2787.25	2527.27	6803.98	2660.45	6654.53	14490.06	4885.59	649.08
Experiment	Fig. 6 (t-SNE, U)	Fig. 6 (t-SNE, S)	Fig. 6 (ResNet, U)	Fig. 6 (ResNet, S)	Fig. 6 (Siamese, U)	Fig. 6 (Siamese, S)	Fig. 6 (S, Exact)	
Time	5373.26	2644.21	6075.75	2332.44	6281.83	30564.37	4893.62	
Experiment	Fig. 8 (synthetic)	Fig. 8 (face)	Fig. 9 (1st row)	Fig. 9 (2nd row)	Fig. 9 (3rd row)			
Time	365.95	4814.96	1181.47	1488.40	1597.77			

4.3. Manifold embedding

QQE can be used for manifold embedding in a lower dimensional embedding space where the embedding distribution can be determined by the user. As an initialization, the high dimensional data are embedded in a lower dimensional embedding space using a dimensionality reduction method. Thereafter, the low dimensional embedding data are transformed to a desired distribution using QQE.

Any dimensionality reduction method can be utilized for the initialization of data in the low dimensional subspace. Some examples are PCA (Ghojogh & Crowley, 2019) (or classical MDS Cox & Cox, 2008), FDA (Ghojogh et al., 2019), Isomap (Tenenbaum et al., 2000), LLE (Roweis & Saul, 2000), t-SNE (Van Der Maaten, 2009), and deep features like triplet Siamese features (Schroff et al., 2015) and ResNet features (He et al., 2016). By initialization, an initial embedding of data is obtained in the low dimensional embedding space.

After the initialization, a reference sample is drawn from the reference distribution or is taken from the user. The dimensionality of the reference sample is equal to the dimensionality of the low dimensional embedding space; in other words, the reference sample is in the low dimensional space. We transform the distribution of the low dimensional data to the reference distribution using QQE. Again, the distribution transformation can be either to the exact or shape of the desired distribution. The proposed methods for distribution transformation to the exact reference distribution or shape of desired distribution were explained in Sections 4.1 and 4.2 and can be used here for distribution transformation in the low dimensional embedding space.

4.4. Unsupervised and supervised embedding

QQE, for both tasks of distribution transformation (see Sections 4.1 and 4.2) and manifold embedding (see Section 4.3), can be used in either supervised or unsupervised manners. In the following, we explain these two cases:

- In the *unsupervised* form, all data points are seen together as a cloud of data and the distribution of all data points is transformed to a desired distribution. The unsupervised QQE algorithm for distribution transformation transforms the entire dataset to have the desired distribution. For manifold embedding, unsupervised QQE initializes the embedding data into the low dimensional space and then transforms the entire embedded data to have the desired distribution.
- In the *supervised* manner, the data points of each class are transformed to have a desired distribution. Hence, in this manner, the user may choose different distributions for each class. The supervised QQE for distribution transformation transforms the distribution of every class to a desired distribution. For manifold learning, supervised QQE initializes the embedding data into the low dimensional space and then transforms the embedded data of every class to a desired distribution. Note that QQE for manifold learning can be supervised regardless of whether the dimensionality reduction method used for initialization is unsupervised or supervised.

It is noteworthy that in both unsupervised and supervised manners of QQE, the distribution transformation and manifold embedding can be either to the exact reference distribution (see Section 4.1) or to the shape of reference distribution (see Section 4.2).

5. Experiments

In this section, we report the experimental results. The code for this paper and its experiments can be found in our Github repository.[1] The hardware used for the experiments was Intel Core-i7 CPU with the base frequency 1.80 GHz and 16 GB RAM. Table 1 reports the timing of different experiments for giving a sense of pacing in QQE algorithm. Note that the time complexity of QQE algorithm is $\mathcal{O}(n^3 + ndk)$ because of the assignment problem (Edmonds & Karp, 1972) and the optimization steps, respectively. Improvement of time complexity of QQE is a possible future direction discussed in Section 6. Note that for all experiments in this article, unless specifically mentioned, we set $\lambda = 0.1$, $\eta = 0.01$, and $k = 10$. A comprehensive discussion on the effect of these hyperparameters will be provided in Section 5.3.

5.1. Quantitative measures used for difference of distributions

In our experimental results, in addition to illustrating the visualization of distribution transformation either in the input space or in the embedding space, we report several quantitative measurements for validating distribution transformation theoretically. Table 2 reports the quantitative measurements for all experiments, showing the improvement of change of distributions to the desired distributions using the QQE algorithm. The three measures used are KL-divergence, MMD and HSIC which are briefly defined below.

Assume we have two samples from the following distributions: $\{\boldsymbol{x}_i\}_{i=1}^n \sim \mathcal{P}$ and $\{\boldsymbol{y}_i\}_{i=1}^n \sim \mathcal{Q}$. The first used measure is KL-divergence (Kullback & Leibler, 1951). The KL-divergence for discrete samples is defined as:

$$\mathrm{KL}(\mathcal{P} \parallel \mathcal{Q}) := \sum_{i=1}^{n} \mathcal{P}(\boldsymbol{x}_i) \log\Big(\frac{\mathcal{P}(\boldsymbol{x}_i)}{\mathcal{Q}(\boldsymbol{y}_i)}\Big), \tag{23}$$

for the difference of distributions $\mathcal{P}$ and $\mathcal{Q}$. We estimate $\mathcal{P}(\boldsymbol{x}_i)$ and $\mathcal{Q}(\boldsymbol{y}_i)$ using kernel density estimation with Gaussian kernels and the Scott's rule (Scott, 2015). Note that $\mathrm{KL} \geq 0$ where $\mathrm{KL} = 0$ means

[1] https://github.com/bghojogh/Quantile-Quantile-Embedding

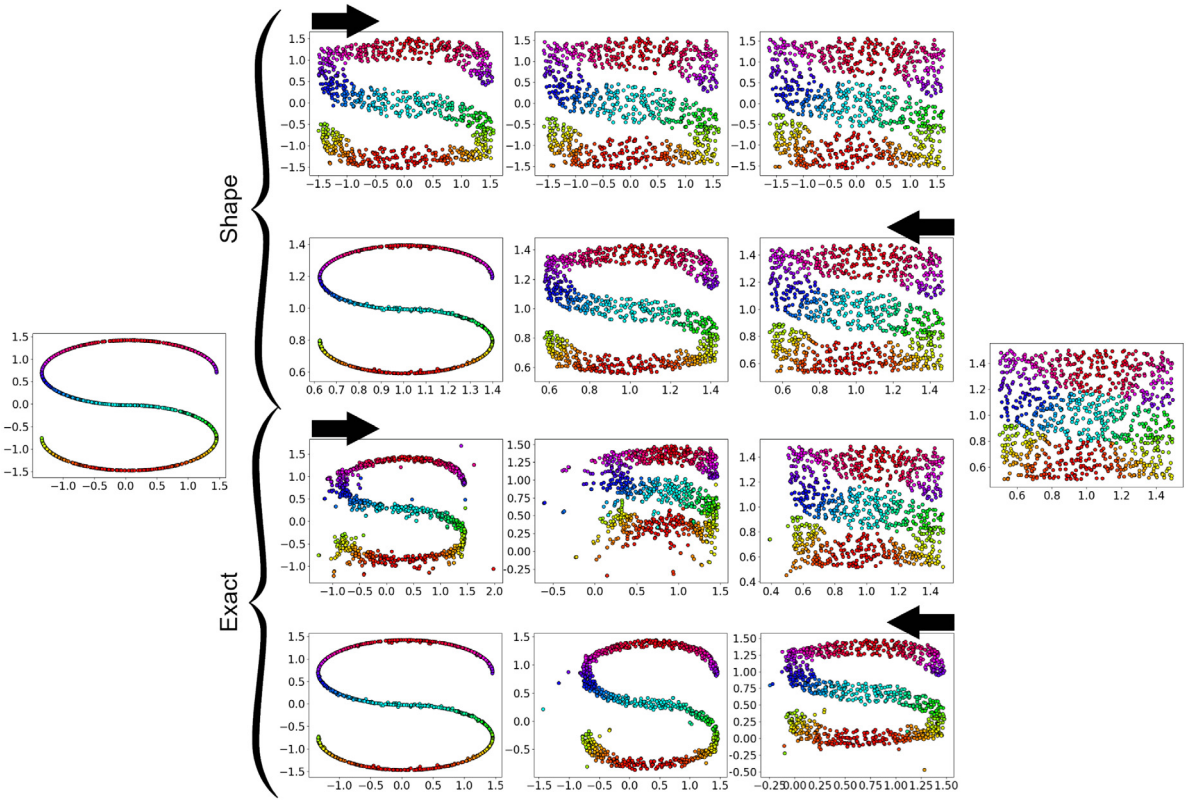

Fig. 1. Distribution transformation of S-shape and uniform data to each other. The first and second pair of rows correspond to transformation of shape and exact distributions, respectively. The arrows show the direction of gradual changes.

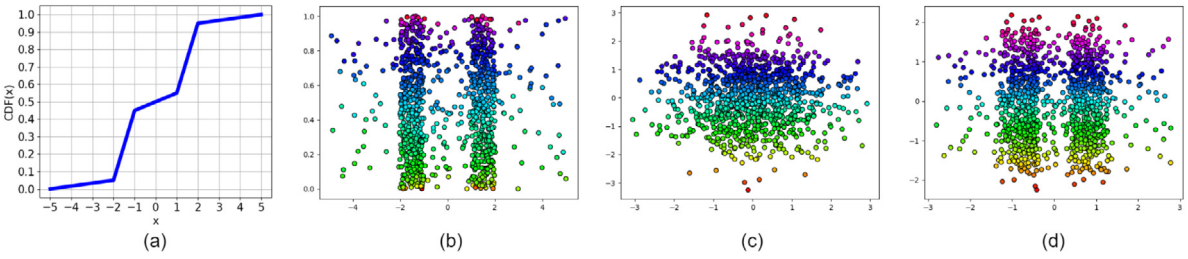

Fig. 2. Distribution transformation using (a) CDF of reference distribution: (b) the reference data, (c) Gaussian data, and (d) transformed data.

the two distributions are equivalent. After applying QQE for distribution transformation or manifold embedding, we would expect the KL-divergence between the sample $\{\boldsymbol{x}_i\}_{i=1}^n$ and the reference sample $\{\boldsymbol{y}_i\}_{i=1}^n$ to be reduced. Note that the amount of reduction of KL-divergence is not necessarily meaningful as KL-divergence does not have any upperbound.

The second measure used for difference of distributions is MMD (Gretton et al., 2007, 2012). It compares the moments of distributions using distances in the feature space (Schölkopf, 2001). Let $\boldsymbol{\phi}(\boldsymbol{x})$ be the pulling function from the input to the feature space and $k(\boldsymbol{x}_i, \boldsymbol{x}_j) := \boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_j)$ be the kernel function (Hofmann et al., 2008). It is defined as:

$$\begin{aligned}\text{MMD}^2(\mathcal{P}, \mathcal{Q}) &:= \Big\| \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{x}_i) - \frac{1}{n}\sum_{i=1}^n \boldsymbol{\phi}(\boldsymbol{y}_i) \Big\|^2 \\ &= \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(\boldsymbol{y}_i, \boldsymbol{y}_j) \\ &\quad - \frac{2}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(\boldsymbol{x}_i, \boldsymbol{y}_j), \end{aligned} \tag{24}$$

where $\|.\|$ denotes a norm in the feature/Hilbert space. Note that MMD ≥ 0 where MMD $= 0$ means the two distributions are equivalent. After applying QQE for distribution transformation or manifold embedding, it is mostly expected to have smaller MMD between the sample $\{\boldsymbol{x}_i\}_{i=1}^n$ and the reference sample $\{\boldsymbol{y}_i\}_{i=1}^n$. As MMD does not have any upper-bound, the amount of reduction of MMD is not important but the reduction itself is mostly expected.

The third method used in this paper for quantitative measurements is HSIC (Gretton et al., 2005). It estimates the dependence of two random variables by computing the correlation of the pulled data $\boldsymbol{\phi}(\boldsymbol{x}_i)$ and $\boldsymbol{\phi}(\boldsymbol{x}_j)$ using the Hilbert–Schmidt norm of their cross-covariance. One can refer to Gubner (2006) for definitions of the Hilbert–Schmidt norm and the cross-covariance matrix of two random variables. An empirical estimate of HSIC between samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$ is (Gretton et al., 2005):

$$\text{HSIC}(X, Y) := \frac{1}{(n-1)^2}\textbf{tr}(\boldsymbol{K}_x \boldsymbol{H} \boldsymbol{K}_y \boldsymbol{H}), \tag{25}$$

where $\textbf{tr}(.)$ denotes the trace of matrix and $\boldsymbol{K}_x$ and $\boldsymbol{K}_y$ are kernels over samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{\boldsymbol{y}_i\}_{i=1}^n$, respectively (Hofmann et al., 2008). $\mathbb{R}^{n\times n} \ni \boldsymbol{H} := \boldsymbol{I} - (1/n)\boldsymbol{1}\boldsymbol{1}^\top$ is the centering matrix where $\boldsymbol{I}$ and $\boldsymbol{1}$ denote the identity matrix and the vector of ones, respectively. Note

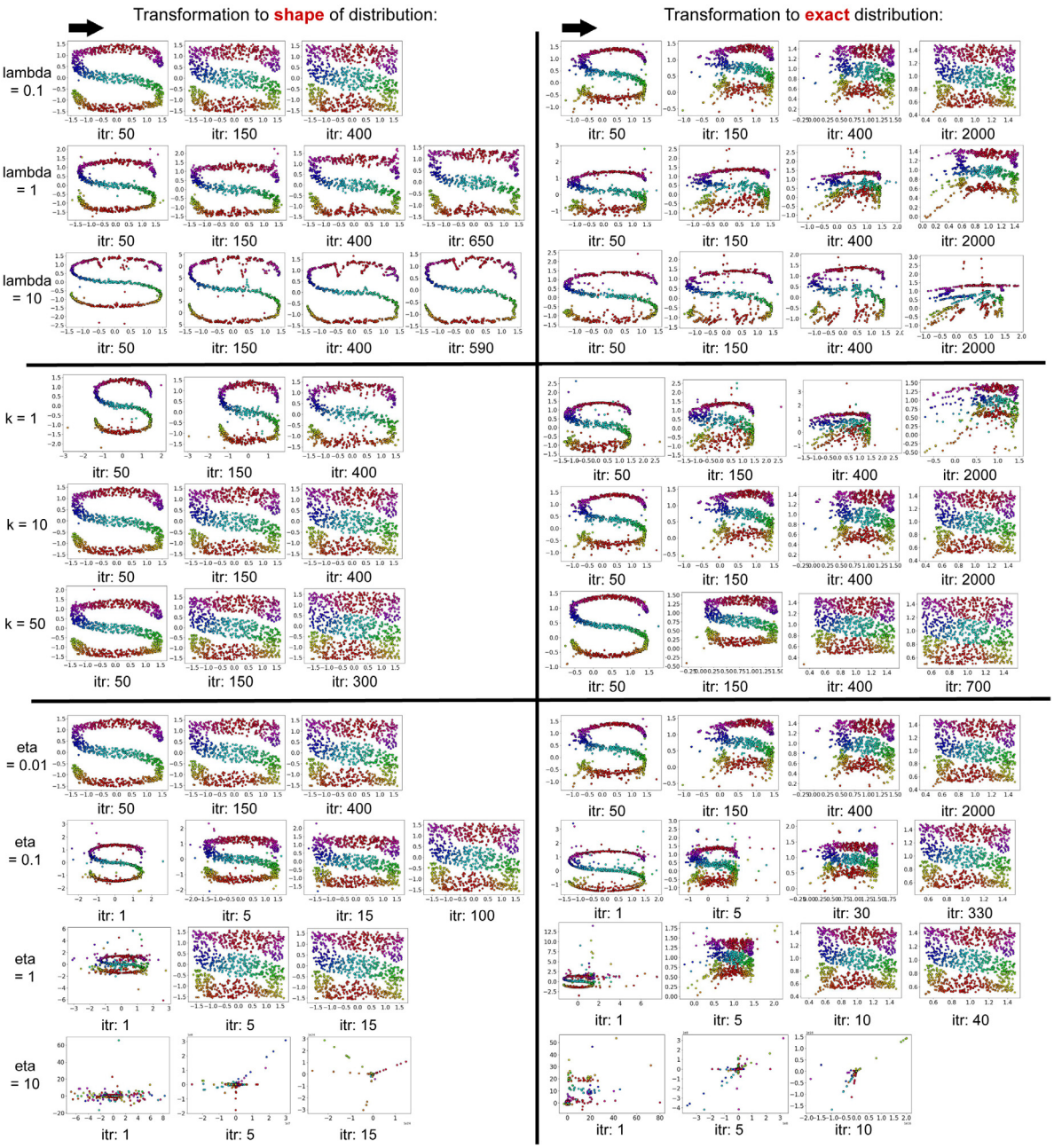

Fig. 3. Analysis of effect of hyperparameters on the performance of QQE for both transformations to the shape of distribution and exact distribution. The initial and reference distributions are as of the left and right distributions depicted in Fig. 1.

that HSIC ≥ 0 where HSIC $= 0$ means the two random variables are independent. The more the HSIC is, the more dependent the variables are. After applying QQE for distribution transformation or manifold embedding, it is mostly expected to have larger HSIC between the sample $\{\boldsymbol{x}_i\}_{i=1}^n$ and the reference sample $\{\boldsymbol{y}_i\}_{i=1}^n$. As HSIC does not have any upperbound, the amount of increase of HSIC is not important but the increase itself is mostly expected. It is noteworthy that the trend of decrease in KL-divergence and MMD often coincide with the trend of increase in HSIC; although in some rare cases, this coincident does not hold.

5.2. Distribution transformation for synthetic data

To visually show how distribution transformation works, we report the results of QQE on some synthetic datasets. In the following, we report several different possible cases for distribution transformation.

Fig. 4. Distribution transformation of facial images (Cambridge, 2020; Samaria & Harter, 1994) without eyeglasses to the shape of images with eyeglasses. The arrow shows the direction of gradual changes.

Table 2
The quantitative evaluation of QQE embeddings for experiments in this paper. Letters U and S denote unsupervised and supervised QQE approaches, respectively. For supervised cases, the reported number is the average of that measure among classes. In every cell of table, the left-side and right-side numbers correspond to before and after applying QQE algorithm, respectively.

Experiment	Fig. 1 (1st row)	Fig. 1 (2nd row)	Fig. 1 (3rd row)	Fig. 1 (4th row)	Fig. 2	Fig. 4		
KL-divergence	4.90E−2 | 3.58E−2	3.91E−2 | 3.70E−2	4.90E−2 | 4.04E−2	3.91E−2 | 4.39E−2	3.18E−1 | 2.56E−1	2.40E−3 | 1.23E−3		
MMD^2	5.99E−1 | 5.84E−1	2.22E−16 | 6.07E−5	5.99E−1 | 5.47E−5	2.22E−16 | 3.86E−1	1.92E−1 | 1.87E−1	1.68E−2 | 1.68E−2		
HSIC	7.33E−5 | 8.11E−5	2.11E−5 | 1.73E−5	7.33E−5 | 1.95E−5	2.11E−5 | 5.68E−5	3.18E−4 | 3.25E−4	8.47E−3 | 8.47E−3		
Experiment	Fig. 5 (PCA, U)	Fig. 5 (PCA, S)	Fig. 5 (FDA, U)	Fig. 5 (FDA, S)	Fig. 5 (Isomap, U)	Fig. 5 (Isomap, S)	Fig. 5 (LLE, U)	Fig. 5 (LLE, S)
KL-divergence	2.10E−1 | 1.05E−1	9.23E−2 | 6.33E−3	1.30E−1 | 7.78E−2	8.22E−2 | 1.42E−2	2.52E−1 | 1.20E−1	8.45E−2 | 1.80E−2	3.26E−1 | 1.83E−1	1.45E−1 | 7.10E−2
MMD^2	6.48E−1 | 5.92E−1	7.46E−1 | 7.46E−1	6.64E−1 | 5.93E−1	9.76E−1 | 9.77E−1	6.86E−1 | 6.21E−1	7.85E−1 | 7.88E−1	6.29E−1 | 7.60E−1	8.12E−1 | 7.79E−1
HSIC	7.52E−5 | 8.89E−5	1.82E−2 | 2.22E−2	1.26E−4 | 1.50E−4	1.65E−2 | 2.03E−2	9.37E−5 | 9.20E−5	1.62E−2 | 1.91E−2	1.28E−4 | 1.62E−4	6.46E−3 | 5.53E−3
Experiment	Fig. 5 (t-SNE, U)	Fig. 5 (t-SNE, S)	Fig. 5 (S, Exact)					
KL-divergence	5.23E−2 | 5.43E−2	7.92E−2 | 2.43E−2	7.98E−2 | 4.34E−2					
MMD^2	8.59E−1 | 8.57E−1	8.65E−1 | 8.62E−1	7.59E−1 | 2.55E−1					
HSIC	1.43E−4 | 1.44E−4	1.05E−3 | 7.14E−4	1.76E−2 | 1.66E−2					
Experiment	Fig. 6 (PCA, U)	Fig. 6 (PCA, S)	Fig. 6 (FDA, U)	Fig. 6 (FDA, S)	Fig. 6 (Isomap, U)	Fig. 6 (Isomap, S)	Fig. 6 (LLE, U)	Fig. 6 (LLE, S)
KL-divergence	2.66E−1 | 1.14E−1	1.56E−1 | 3.77E−2	1.93E−1 | 8.36E−2	1.61E−1 | 4.41E−2	2.42E−1 | 1.02E−1	1.58E−1 | 3.85E−2	6.50E−1 | 5.51E−1	1.50E−1 | 1.19E−1
MMD^2	4.48E−1 | 4.81E−1	5.63E−1 | 5.46E−1	4.20E−1 | 4.11E−1	7.81E−1 | 7.78E−1	8.53E−1 | 8.53E−1	5.43E−1 | 5.45E−1	3.30E−1 | 2.06E−1	1.23E0 | 1.24E0
HSIC	4.62E−5 | 4.70E−5	1.45E−2 | 1.91E−2	3.11E−5 | 3.16E−5	4.46E−2 | 6.01E−2	1.47E−5 | 1.47E−5	1.95E−3 | 2.49E−3	5.54E−5 | 6.52E−5	1.40E−2 | 1.64E−2
Experiment	Fig. 6 (t-SNE, U)	Fig. 6 (t-SNE, S)	Fig. 6 (ResNet, U)	Fig. 6 (ResNet, S)	Fig. 6 (Siamese, U)	Fig. 6 (Siamese, S)	Fig. 6 (S, Exact)	
KL-divergence	5.11E−2 | 4.42E−2	1.10E−1 | 4.22E−2	1.89E−1 | 8.83E−2	1.65E−1 | 3.88E−2	1.02E−1 | 5.66E−2	1.31E−1 | 3.16E−2	1.48E−1 | 1.03E−1	
MMD^2	8.13E−1 | 8.12E−1	5.49E−1 | 5.48E−1	6.15E−1 | 6.34E−1	6.72E−1 | 6.58E−1	4.87E−2 | 4.85E−2	1.23EE0 | 1.24E0	5.61E−1 | 4.66E−1	
HSIC	1.87E−5 | 1.85E−5	2.62E−3 | 3.24E−3	2.28E−5 | 2.32E−5	4.29E−2 | 5.79E−2	5.30E−5 | 5.49E−5	2.10E−2 | 2.31E−2	1.46E−2 | 3.65E−2	
Experiment	Fig. 8 (synthetic)	Fig. 8 (face)	Fig. 9 (1st row)	Fig. 9 (2nd row)	Fig. 9 (3rd row)			
KL-divergence	2.00E−2 | 3.13E−2	1.27E−3 | 1.11E−3	3.35E−12 | 1.54E−14	4.33E−15 | 3.29E−15	3.33E−15 | 2.93E−15			
MMD^2	8.54E−1 | 3.02E−1	1.23E−2 | 1.23E−2	3.61E−1 | 3.04E−1	3.33E−1 | 3.13E−1	3.10E−1 | 1.04E−1			
HSIC	4.33E−2 | 4.80E−2	6.02E−3 | 6.02E−3	2.53E−5 | 6.72E−5	1.36E−4 | 3.25E−4	3.21E−4 | 5.46E−4			

Table 3
The average time of QQE iterations for experiments on the impact of hyperparameters, illustrated in Fig. 3. The reported average times are in seconds.

Transformation type	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 10$	$k = 1$	$k = 10$	$k = 50$	$\eta = 0.01$	$\eta = 0.1$	$\eta = 1$	$\eta = 10$
Shape	0.63	0.56	0.61	0.47	0.63	1.11	0.63	0.51	0.49	0.55
Exact	0.61	0.20	0.35	0.35	0.61	1.03	0.61	0.53	0.53	0.48

5.2.1. Standard reference distributions

A simple option for the reference distribution is a standard probability distribution. As an example, we drew a sample of size 1000 from the two dimensional uniform distribution in range $[0.5, 1.5]$ in both dimensions. This sample is depicted at the right hand side of Fig. 1. We also created an S-shape dataset, with mean zero and in range $[-1.5, 1.5]$ in both dimensions, illustrated at the left hand side of Fig. 1. As this figure shows, in transforming the S-shape data to the shape of uniform distribution, the dataset gradually expands to fill the gaps and become similar to the uniform distribution without changing its mean and scale. In transforming to the exact uniform distribution, however, the mean and scale of data change gradually, by translation and contraction, to match the moments of the reference distribution. The runtime for this experiment is reported in Table 1. The KL-divergence, MMD, and HSIC of distribution transformation of S-shape data to either the shape of uniform distribution or the exact uniform distribution are reported in Table 2. As expected, after applying QQE, the KL-divergence and MMD have decreased and HSIC has often increased. Note that transformations to exact distribution mostly have smaller KL-divergence and MMD and larger HSIC compared to transformations to the shape of reference distribution. This is because exact transformation matches all moments while some moments are not matched in shape transformation.

5.2.2. Given reference sample

We can also define the target distribution to transform to using an empirical reference sample. An example is the S-shape data shown in Fig. 1 where we transform the uniform data to its distribution. In shape transformation, two gaps appear first to imitate the S shape and then the stems become narrower iteratively. In exact transformation, however, the mean and scale of data also change. Note that exact transformation is harder than shape transformation because of the change of moments; thus, some points jump at initial iterations and then converge gradually. In Section 6, we report on future work to make QQE more robust to these jumps. The runtime for this experiment is reported in Table 1. The KL-divergence, MMD, and HSIC of distribution transformation of uniform data to either the shape of S-shape distribution or the exact S-shape distribution are reported in Table 2. As expected, after applying QQE, the KL-divergence has decreased and HSIC has often increased. Again, transformation to exact distribution mostly have smaller KL-divergence and MMD and larger HSIC compared to transformation to the shape of reference distribution, for the reason explained before. The experiments of distribution transformation to a standard or a given distribution show that the proposed QQE can change the distribution of data to any distribution. This desired reference distribution can be a simple or a complicated distribution. Moreover, it can be either a theoretical distribution or an available reference sample.

5.2.3. Given cumulative distribution function

Instead of a standard reference distribution or a reference sample, the user can give a desired CDF for the distribution to have. The reference sample can be sampled using the inverse CDF (Ghojogh, Nekoei, Ghojogh, Karray, & Crowley, 2020). The CDF can be multivariate; however, for the sake of visualization, Fig. 2-a shows an example multi-modal univariate CDF. We used this CDF and uniform distribution for the first and second dimensions of the reference sample, respectively, shown in Fig. 2-b. QQE was applied on the Gaussian data shown in Fig. 2-c and its distribution changed to have a CDF similar to the reference CDF (see Fig. 2-d). The runtime for this experiment is reported in Table 1. The KL-divergence, MMD, and HSIC of distribution transformation of Gaussian data to either the given CDF are reported in Table 2. As expected, after applying QQE, the KL-divergence and MMD have decreased and HSIC has increased. This experiment shows that the proposed QQE gives flexibility to user to even choose the desired distribution by a CDF function or plot. This validates the user-friendliness of QQE algorithm.

5.3. Discussion on impact of hyperparameters

Here, we discuss the impact of hyperparameters λ, η, and k on the performance of QQE. QQE is not yet applicable on out-of-sample data (see Section 6) so these parameters cannot be determined by validation;

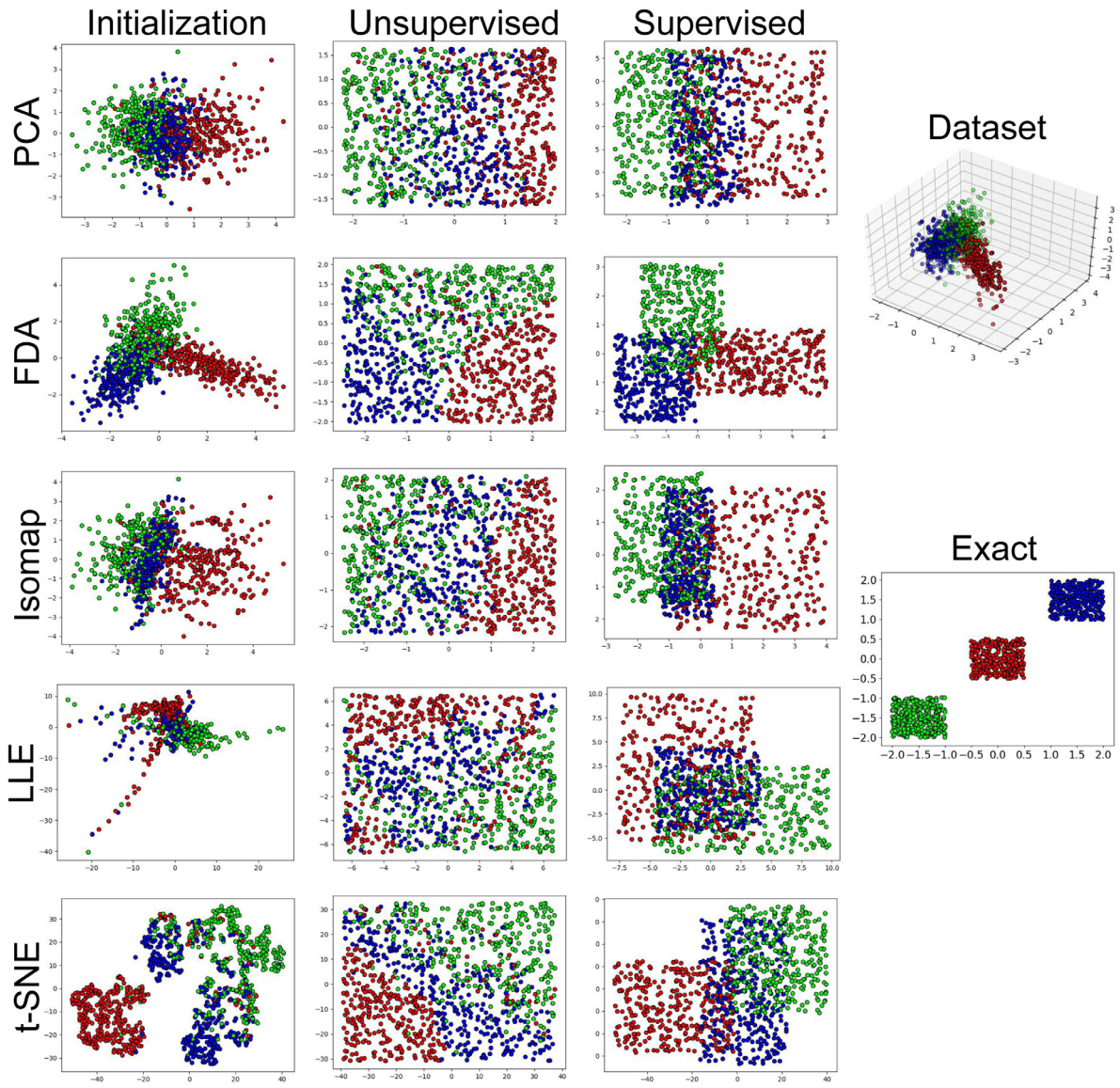

Fig. 5. Unsupervised and supervised exact manifold embedding of the synthetic data with different initializations. Transformation to exact reference distribution is also shown. The initialization of LLE is scaled by constant to be in range of other embeddings.

however, here, we briefly discuss the impact of these hyperparameters. For better understanding of discussion, we illustrate the performance of QQE under different hyperparameter settings in Fig. 3. This illustration shows the impact of hyperparameters on distribution transformation or manifold embedding by QQE if the transformations are performed in the input space or embedding space, respectively. The average time of quasi-Newton iterations in QQE for experiments of Fig. 3 are reported in Table 3.

The regularization parameter λ determines the importance of distance preserving compared to the quantile–quantile plot of distributions. The larger this parameter gets, the less important the distribution transformation becomes compared to preserving distances; hence, the slower the progress of optimization gets. As Fig. 3 illustrates, small enough λ converges both faster and better. The value $\lambda = 0.1$ was empirically found to be proper for different datasets. The learning rate η should be set small enough to have progress in optimization without oscillating behavior. As shown in Fig. 3, larger η makes convergence faster but may result in divergence of optimization. We empirically found $\eta = 0.01$ or $\eta = 0.1$ to be good for different datasets. The larger number of neighbors k results in slower pacing of optimization because of Eqs. (15) and (16). This can be validated by average time of QQE for large value of k reported in Table 3. Very small k, however, does not capture the local patterns of data (Saul & Roweis, 2003). For this reason, as Fig. 3 depicts, small k does not perform perfectly for QQE. The value $k = 10$ is fairly proper for different datasets.

5.4. Distribution transformation for image data

The distribution transformation can be used for any real data such as images. We divided the ORL facial images (Cambridge, 2020; Samaria & Harter, 1994) into two sets of with and without eyeglasses. The set with eyeglasses was taken as the reference sample and we transformed the set without glasses to have the shape of reference distribution. Fig. 4 illustrates the gradual change of two example faces from not having eyeglasses to having them. The glasses have appeared gradually in the eye regions of faces. The runtime for this experiment is reported in Table 1. The KL-divergence, MMD, and HSIC of distribution transformation of facial image data are reported in Table 2. As expected, after applying QQE, the KL-divergence has decreased. As shown by this experiment, the proposed QQE can be useful for image processing and image modification purposes where the distribution of image is

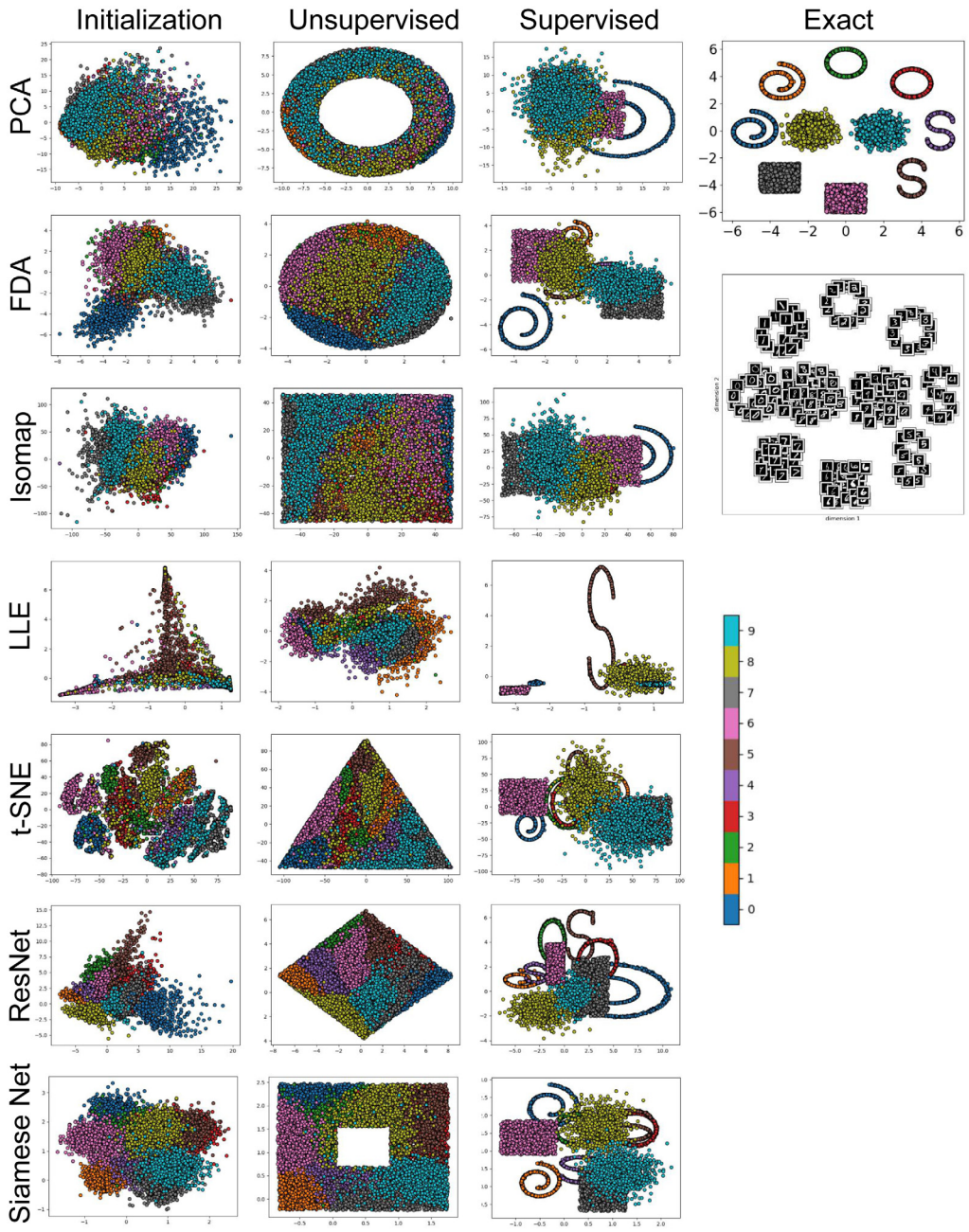

Fig. 6. Unsupervised and supervised exact manifold embedding of the image data with different initializations. Transformation to exact reference distribution is also shown. The initialization of LLE is scaled by constant to be in range of other embeddings.

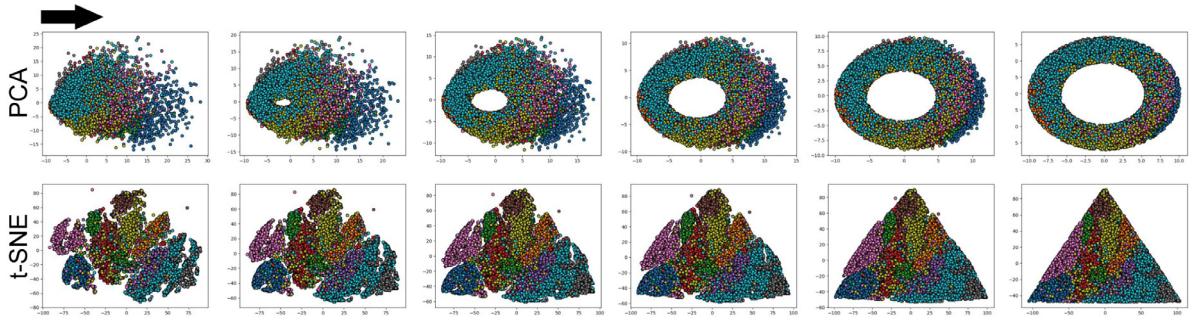

Fig. 7. Some iterations of unsupervised manifold embedding initialized by PCA and t-SNE. The arrow shows the direction of gradual changes.

Table 4
The Recall@k measure for $k \in \{1, 2, 4, 8\}$ for evaluation of separation of classes using QQE algorithm. This table is a quantitative measure of the steps of the experiments shown in Fig. 8.

Synthetic Data	Step 1	Step 2	Step 3	Step 4
Recall@1	71.00	83.20	97.00	100.00
Recall@2	83.60	91.70	98.60	100.00
Recall@4	91.60	96.40	98.80	100.00
Recall@8	95.50	98.70	99.00	100.00
Face (Eye-glasses) Data	Step 1	Step 2	Step 3	Step 4
Recall@1	89.25	97.25	100.00	100.00
Recall@2	95.75	98.75	100.00	100.00
Recall@4	98.75	99.25	100.00	100.00
Recall@8	99.75	99.50	100.00	100.00

changed to a desired theoretical distribution or the distribution of another set of images.

5.5. Manifold embedding for synthetic data

To test QQE for manifold embedding, we created a three dimensional synthetic dataset having three classes shown in Fig. 5. Different dimensionality reduction methods, including PCA (Ghojogh & Crowley, 2019), FDA (Ghojogh et al., 2019), Isomap (Tenenbaum et al., 2000), LLE (Roweis & Saul, 2000), and t-SNE (Van Der Maaten, 2009), were used for initialization (see the first column in Fig. 5). There are multiple experiments shown in Fig. 5 which we explain in the following:

- For our *unsupervised* experiment, we used a uniform distribution as reference and transformed the entire embedded data in an unsupervised manner. As the second column in Fig. 5 shows, the embeddings of the entire dataset have changed to have the *shape* of the uniform distribution but the order and adjacency of classes/points differ depending on the initialization method.
- The results of our *supervised* experiments are shown in the third column in Fig. 5. The desired reference distribution for every class was a uniform distribution and we desired the *shape* of a uniform distribution. As the figure depicts, the supervised QQE has made the shape of distribution of every class uniform without changing its mean and scale.
- The last column of Fig. 5 shows the *supervised* transformation of every embedded class to an *exact* reference distribution. The three exact reference distributions (one for each class) are uniform distributions with different means. In exact transformation, the adjacency of points differ depending on the initialization method but the data patterns are similar so we show only one result.

The runtime for these experiments are reported in Table 1. The KL-divergence, MMD, and HSIC of manifold embedding by QQE are reported in Table 2. As expected, after applying QQE, the KL-divergence and MMD have often decreased and HSIC has often increased.

5.6. Image manifold embedding

QQE can be used for manifold embedding of real data such as images. For the experiments, we sampled 10000 images from the MNIST digit dataset (LeCun, Bottou, Bengio, & Haffner, 1998) with 1000 images per digit. This sampling is because of computational reasons for the time complexity of QQE (see Section 6). We used different initialization methods, i.e., PCA (Ghojogh & Crowley, 2019), FDA (Ghojogh et al., 2019), Isomap (Tenenbaum et al., 2000), LLE (Roweis & Saul, 2000), t-SNE (Van Der Maaten, 2009), ResNet-18 features (He et al., 2016) (with cross entropy loss after the embedding layer), and deep triplet Siamese features (Schroff et al., 2015) (with ResNet-18 as the backbone network). Any embedding space dimensionality can be used but here, for visualization, we took it to be two. The initialized embeddings are illustrated in the first column in Fig. 6.

Fig. 6 shows the results of experiments which we explain in the following:

- For *unsupervised* QQE, we took ring stripe, filled circle, uniform (square), Gaussian mixture model, triangle, diamond, and thick square as the reference distribution for embedding initialized by PCA, FDA, Isomap, LLE, t-SNE, ResNet, and Siamese net, respectively. As shown in the second column in Fig. 6, the shape of entire embedding has changed to the desired while the local distances are preserved as much as possible. Fig. 7 illustrates some iterations of changes in PCA and t-SNE embeddings as examples.
- For *supervised* transformation to the *shape* of references distributions, we used different distributions to show that QQE can use any various references for different classes. Helix, circle, S-shape, uniform, and Gaussian were used for the digits 0/1, 2/3, 4/5, 6/7, 8/9, respectively. The third column in Fig. 6 depicts the supervised transformation to shapes of distributions.
- The fourth column in Fig. 6 shows the *supervised* QQE embedding to the *exact* reference distributions. We set the means of reference distributions to be on a global circular pattern. As the fourth column in Fig. 6 shows, it resulted in the transformation of classes to the exact reference distributions on a circular pattern. The images of embedded digits are also shown in this figure.

The runtime for these experiments are reported in Table 1. The KL-divergence, MMD, and HSIC of manifold embedding by QQE are reported in Table 2. As expected, after applying QQE, the KL-divergence and MMD have often decreased and HSIC has often increased. The experiments of manifold embedding for both synthetic and image data show that the proposed QQE fills the gap of having a manifold learning method with ability to choose the embedding distribution. This is important because the manifold learning methods so far did not give this freedom to user or they forced a specific distribution.

5.7. QQE for separation of classes

QQE can be used for separation and discrimination of classes; although, it does not yet support out-of-sample data (see Section 6). For this, reference distributions with far-away means can be chosen where transformation to the exact distribution is used. Hence, the classes move away to match the first moments of reference distributions. We experimented this for both synthetic and image data. A two dimensional synthetic dataset with three mixed classes was created as shown in Fig. 8. The three classes are gradually separated by QQE to match three Gaussian reference distributions with apart means.

For image data, we used the ORL face dataset (Samaria & Harter, 1994) with two classes of faces with and without eyeglasses. The distribution transformation was performed in the input (pixel) space. The two dimensional embeddings, for visualization in Fig. 8, were obtained using the Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, & Melville, 2018). The dataset was standardized and the reference distributions were set to be two Gaussian distributions with apart means. As the figure shows, the two classes are mixed first but gradually the two classes are completely separated by QQE.

The runtime for both of these experiments are reported in Table 1. The KL-divergence, MMD, and HSIC of separation of classes for both synthetic and image data by QQE are reported in Table 2. As expected, after applying QQE, the KL-divergence and MMD have often decreased and HSIC has often increased. Furthermore, the quantitative evaluation of the separation of classes using QQE is reported in Table 4 for both synthetic and image data. Following the literature (Nguyen & De Baets, 2020; Qian, Shang, Sun, Hu, Li, & Jin, 2019; Sikaroudi, Ghojogh, Karray, Crowley, & Tizhoosh, 2020), we used the Recall@k measure for supervised evaluation of embedding in terms of discrimination of

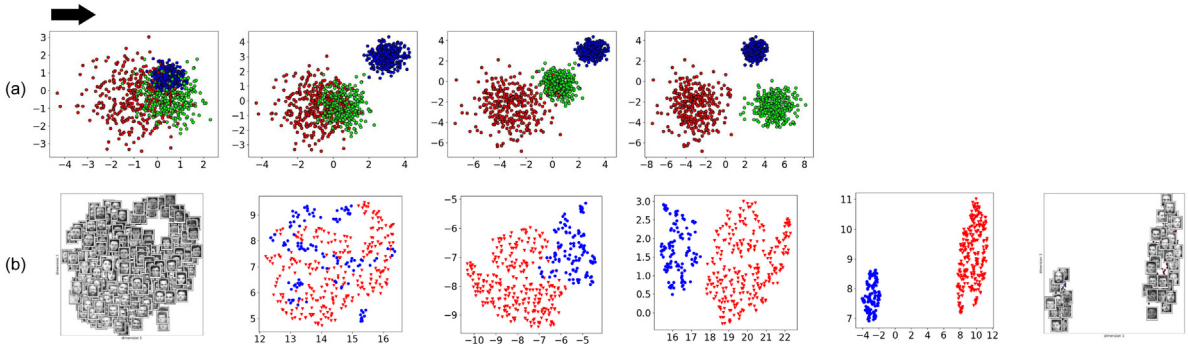

Fig. 8. Separation and discrimination of classes in synthetic and image data. The arrow shows the direction of gradual changes.

classes. As this table shows, QQE has improved the separation of classes by its steps. This improvement in separation of classes can also be seen in Fig. 8. This experiment shows that the proposed QQE can be useful for separation and discrimination of classes either in the input space or embedding space. Discrimination of classes helps better classification of data as well as better representation of data in terms of classes.

5.8. Evaluating QQE for histopathology data

Although the MNIST and ORL face datasets are also real-world datasets, we evaluated the proposed QQE on medical image data as another real dataset. Finding an informative embedding space for extracting features from medical images is useful for image search and finding similar tumorous image patches in hospital's archives (Kalra et al., 2020); therefore, it can be used for automatic cancer diagnosis. We used the Colorectal Cancer (CRC) dataset (Kather et al., 2016) which contains (150×150)-pixel image patches from colorectal histopathology whole slide images. This dataset contains eight tissue types which are background (empty), adipose tissue, mucosal glands, debris, immune cells (lymphoma), complex stroma, simple stroma, and tumor epithelium. Some examples of these tissue types are shown in Fig. 9. We applied QQE manifold embedding on this dataset where Siamese network (Schroff et al., 2015) with the Fisher Discriminant Triplet (FDT) loss (Ghojogh, Sikaroudi, et al., 2020) and ResNet-18 backbone (He et al., 2016) was employed for initialization. For this experiment, $\eta = 0.1$ was found to be proper without oscillation. We set the embedding dimensionality to be 128 to show that QQE also works well on multi-dimensional embedding spaces in addition to two-dimensional spaces.

Fig. 9 illustrates some iterations of applying QQE on the histopathology embedding. We used UMAP (McInnes et al., 2018) for 2D visualizations of 128-dimensional embeddings. The runtime for these experiments are reported in Table 1. The KL-divergence, MMD, and HSIC of these experiments are reported in Table 2. For unsupervised shape transformation, we used a multivariate Gaussian reference distribution. As shown in the figure, the UMAP of entire embedding becomes like Gaussian because the embedding is transformed to be Gaussian. For supervised shape transformation, we considered a multivariate Gaussian distribution for each tissue type. As expected, the UMAP visualization of embedding shows that each class has a Gaussian form eventually. We also performed experiment on supervised transformation to exact reference distribution. For this, we considered eight multivariate Gaussian distributions placed on a global circular pattern similar to what we had in Fig. 6. The UMAP visualization of transformed embedding validates that the tissue embeddings have been placed on a global circular pattern using QQE.

The distribution transformation and manifold embedding for histopathology data have various applications and usages. Although deep metric learning has extracted useful features for tissue types, the embedding of patches have not been separated completely. Both supervised exact transformation and supervised shape transformation can be used to separate the tissue types as desired in the embedding space. In the former, the relative locations of tissue types in the embedding space are also desired to be chosen by user. However, in the latter, the distribution of every tissue type is noticed without changing the relative locations of tissue types in the space. Therefore, QQE may be used for discriminating tumorous tissues from the normal tissues for better cancer diagnosis. The unsupervised transformation can also be used to change the distribution of all tissue types together in the embedding space. These transformations and embeddings can be used in hospitals for several reasons. One possible reason may be that the classifier model, to which the embeddings of tissues are going to be fed, requires a specific distribution to work better. Another reason can be the request of doctors and specialists to analyze tissue types in specific distributions. For any reason which requires distribution transformation or manifold embedding with ability to choose the embedding distribution, QQE can be useful in practice.

6. Conclusion and future directions

In this paper, we proposed QQE for distribution transformation and manifold embedding. This method can be used for both transforming to the exact reference distribution or its shape. Both unsupervised and supervised versions of this method were also proposed. The proposed method was based on quantile–quantile plot which is usually used in visual statistical tests. Experiments were performed on synthetic data, facial images, digit images, and medical histopathology images. We showed that QQE can be used for transforming distribution of data to any desired simple or complicated distribution. The desired distribution can be a theoretical PDF/CDF or an available reference sample. Experiments showed QQE can also be used for modifying images to have a specific distribution or the distribution of another set of images. We also showed that QQE can be used for separation of classes in the input space or embedding space. Experiments on medical data demonstrated that QQE can be useful for practical purposes such as discriminating tumorous tissues from normal ones for better cancer diagnosis.

There exist several possible future directions. The first future direction is to improve the time complexity of QQE. Since the complexity of QQE is $\mathcal{O}(n^3)$, dealing with big data would be a challenge for this initial version. Thus, the immediate future direction for research would be to develop a more sample-efficient approach including handling large datasets. Handling out-of-sample data is another possible future direction. Moreover, QQE uses the least squares problem which is not very robust. Because of this, especially if the moments of data and reference distribution differ significantly and we want to transform to the exact reference distribution, some jumps of some data points may happen at initial iterations. This results in later convergence of QQE. One may investigate high breakdown estimators for robust regression (Yohai, 1987) to make QQE more robust and faster.

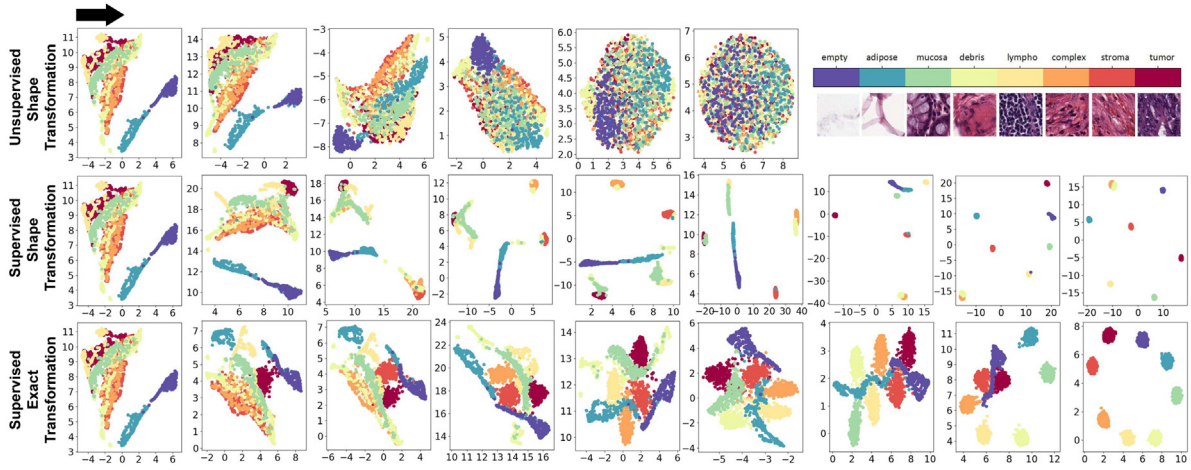

Fig. 9. Applying QQE algorithm, for unsupervised shape transformation, supervised shape transformation, and supervised exact transformation, on colorectal histopathology embedding. The initialization method for manifold embedding was FDT loss with embedding dimensionality of 128. UMAP is used for 2D visualization of embeddings.

CRediT authorship contribution statement

Benyamin Ghojogh: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Writing - Original Draft, Writing - Review & Editing. **Fakhri Karray:** Supervision, Writing - Original Draft, Writing - Review & Editing. **Mark Crowley:** Supervision, Writing - Original Draft, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by a Discovery Grant by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2018-04381).

Appendix A. Proof of Proposition 1

Consider the first part of the cost function:

$$\mathcal{L}_1 := \frac{1}{2}\sum_{i=1}^{n}\|\boldsymbol{x}_i - \boldsymbol{y}_{\sigma(i)}\|_2^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{l=1}^{d}(x_{i,l} - y_{\sigma(i),l})^2 \implies \frac{\partial \mathcal{L}_1}{\partial x_{i,l}} = (x_{i,l} - y_{\sigma(i),l}).$$

Consider the second part of the cost function:

$$\mathcal{L}_2 := \frac{1}{2a}\sum_{i=1}^{n}\sum_{j\in\mathcal{N}_i}\frac{\big(d_x(i,j) - d_x^0(i,j)\big)^2}{d_x^0(i,j)}.$$

By chain rule, $\partial\mathcal{L}_2/\partial x_{i,l} = \partial\mathcal{L}_2/\partial d_x(i,j) \times \partial d_x(i,j)/\partial x_{i,l}$. The first derivative is:

$$\frac{\partial \mathcal{L}_2}{\partial d_x(i,j)} = \frac{1}{a}\sum_{j\in\mathcal{N}_i}\frac{d_x(i,j) - d_x^0(i,j)}{d_x^0(i,j)},$$

and using the chain rule, the second derivative is $\partial d_x(i,j)/\partial x_{i,l} = \partial d_x(i,j)/\partial d_x^2(i,j) \times \partial d_x^2(i,j)/\partial x_{i,l}$. We have:

$$\frac{\partial d_x(i,j)}{\partial d_x^2(i,j)} = 1/\frac{\partial d_x^2(i,j)}{\partial d_x(i,j)} = 1/(2\,d_x(i,j)).$$

$$d_x^2(i,j) = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 = \sum_{k=1}^{p}(x_{i,l} - x_{j,l})^2.$$

$$\frac{\partial d_x^2(i,j)}{\partial x_{i,l}} = 2\,(x_{i,l} - x_{j,l}), \quad \therefore \frac{\partial d_x(i,j)}{\partial x_{i,l}} = \frac{x_{i,l} - x_{j,l}}{d_x(i,j)}. \tag{26}$$

$$\therefore \quad \frac{\partial \mathcal{L}_2}{\partial x_{i,l}} = \frac{1}{a}\sum_{j\in\mathcal{N}_i}\frac{d_x(i,j) - d_x^0(i,j)}{d_x(i,j)\,d_x^0(i,j)}(x_{i,l} - x_{j,l}).$$

Considering both parts of the cost function, the gradient is as in the proposition. □

Appendix B. Proof of Proposition 2

The second derivative is the derivative of the first derivative, i.e., Eq. (15). Hence:

$$\frac{\partial^2 \mathcal{L}}{\partial x_{i,l}^2} = 1 + \frac{\lambda}{a}\sum_{j\in\mathcal{N}_i}\frac{\partial}{\partial x_{i,l}}\Big(\frac{d_x(i,j) - d_x^0(i,j)}{d_x(i,j)\,d_x^0(i,j)}(x_{i,l} - x_{j,l})\Big).$$

$$\begin{aligned}
&\frac{\partial}{\partial x_{i,l}}\Big(\frac{d_x(i,j) - d_x^0(i,j)}{d_x(i,j)\,d_x^0(i,j)}(x_{i,l} - x_{j,l})\Big) \\
&\quad = (x_{i,l} - x_{j,l})\frac{\partial}{\partial x_{i,l}}\Big(\frac{d_x(i,j) - d_x^0(i,j)}{d_x(i,j)\,d_x^0(i,j)}\Big) \\
&\qquad + \frac{d_x(i,j) - d_x^0(i,j)}{d_x(i,j)\,d_x^0(i,j)}\underbrace{\frac{\partial}{\partial x_{i,l}}(x_{i,l} - x_{j,l})}_{=1}.
\end{aligned}$$

$$\begin{aligned}
\frac{\partial}{\partial x_{i,l}}\Big(\frac{d_x(i,j) - d_x^0(i,j)}{d_x(i,j)\,d_x^0(i,j)}\Big) &= \frac{1}{d_x^0(i,j)}\frac{\partial}{\partial x_{i,l}}\Big(1 - \frac{d_x^0(i,j)}{d_x(i,j)}\Big) \\
&= \frac{1}{d_x^0(i,j)}\underbrace{\frac{\partial}{\partial x_{i,l}}(1)}_{=0} - \underbrace{\frac{d_x^0(i,j)}{d_x^0(i,j)}}_{=1}\frac{\partial}{\partial x_{i,l}}\Big(\frac{1}{d_x(i,j)}\Big) \\
&= \frac{1}{d_x^2(i,j)}\frac{\partial}{\partial x_{i,l}}(d_x(i,j)) \overset{(26)}{=} \frac{(x_{i,l} - x_{j,l})}{d_x^3(i,j)}.
\end{aligned}$$

Putting all parts of derivative together gives the second derivative. □

References

Alipanahi, B., & Ghodsi, A. (2011). Guided locally linear embedding. *Pattern Recognition Letters*, *32*(7), 1029–1035.

Allen, H. (1914). The storage to be provided in impounding reservoirs for municipal water supply. *Transactions of the American Society of Civil Engineers*, *77*, 1539–1669.

Barshan, E., Ghodsi, A., Azimifar, Z., & Jahromi, M. Z. (2011). Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*, *44*(7), 1357–1371.

Cambridge, A. L. (2020). AT&T laboratories cambridge. http://cam-orl.co.uk/facedatabase.html. (Accessed 01 January 2020).

Chaudhuri, P. (1996). On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association*, *91*(434), 862–872.

Cox, M. A., & Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization* (pp. 315–347). Springer.

Dhar, S. S., Chakraborty, B., & Chaudhuri, P. (2014). Comparison of multivariate distributions using quantile–quantile plots and related tests. *Bernoulli*, *20*(3), 1484–1506.

Easton, G. S., & McCulloch, R. E. (1990). A multivariate generalization of quantile-quantile plots. *Journal of the American Statistical Association*, *85*(410), 376–386.

Edmonds, J., & Karp, R. M. (1972). Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM*, *19*(2), 248–264.

Ferguson, T. S. (1967). *Mathematical statistics: A decision theoretic approach*. Academic Press.

Galton, F. (1874). On a proposed statistical scale. *Nature*, *9*(227), 342.

Galton, F., Foxwell, P., Martin, J. B., Walker, F., Marshall, P., Longstaff, G., et al. (1885). The application of a graphic method to fallible measures [with discussion]. *Journal of the Statistical Society of London*, 262–271.

Ghojogh, B., & Crowley, M. (2019). Unsupervised and supervised principal component analysis: Tutorial. arXiv preprint arXiv:1906.03148.

Ghojogh, B., Karray, F., & Crowley, M. (2019). Fisher and kernel Fisher discriminant analysis: Tutorial. arXiv preprint arXiv:1906.09436.

Ghojogh, B., Nekoei, H., Ghojogh, A., Karray, F., & Crowley, M. (2020). Sampling algorithms, from survey sampling to Monte Carlo methods: Tutorial and literature review. arXiv preprint arXiv:2011.00901.

Ghojogh, B., Sikaroudi, M., Shafiei, S., Tizhoosh, H. R., Karray, F., & Crowley, M. (2020). Fisher discriminant triplet and contrastive losses for training siamese networks. In *2020 international joint conference on neural networks* (pp. 1–7). IEEE.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press Cambridge.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. J. (2007). A kernel method for the two-sample-problem. In *Advances in neural information processing systems* (pp. 513–520).

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*(Mar), 723–773.

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory* (pp. 63–77). Springer.

Gubner, J. A. (2006). *Probability and random processes for electrical and computer engineers*. Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). http://dx.doi.org/10.1109/CVPR.2016.90.

Hinton, G. E., & Roweis, S. T. (2003). Stochastic neighbor embedding. In *Advances in neural information processing systems* (pp. 857–864).

Hofmann, T., Schölkopf, B., & Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 1171–1220.

Hyndman, R. J., & Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, *50*(4), 361–365.

Kalra, S., Tizhoosh, H. R., Shah, S., Choi, C., Damaskinos, S., Safarpoor, A., et al. (2020). Pan-cancer diagnostic consensus through searching archival histopathology images using artificial intelligence. *NPJ Digital Medicine*, *3*(1), 1–15.

Kather, J. N., Weis, C. A., Bianconi, F., Melchers, S. M., Schad, L. R., Gaiser, T., et al. (2016). Multi-class texture analysis in colorectal cancer histology. *Scientific Reports*, *6*(1), 1–11.

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

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Lee, J. A., & Verleysen, M. (2007). *Nonlinear dimensionality reduction*. Springer Science & Business Media.

Leon Harter, H. (1984). Another look at plotting positions. *Communications in Statistics. Theory and Methods*, *13*(13), 1613–1633.

Li, Y., Swersky, K., & Zemel, R. (2015). Generative moment matching networks. In *International conference on machine learning* (pp. 1718–1727).

Loy, A., Follett, L., & Hofmann, H. (2016). Variations of Q–Q Plots: The power of our eyes! *The American Statistician*, *70*(2), 202–214.

Maaten, L. V. D., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*(Nov), 2579–2605.

Marden, J. I. (2004). Positions and QQ plots. *Statistical Science*, *19*(4), 606–614.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Möttönen, J., & Oja, H. (1995). Multivariate spatial sign and rank methods. *Journal of Nonparametric Statistics*, *5*(2), 201–213.

Nguyen, B., & De Baets, B. (2020). Improved deep embedding learning based on stochastic symmetric triplet loss and local sampling. *Neurocomputing*, *402*, 209–219.

Nocedal, J., & Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Oldford, R. W. (2016). Self-calibrating quantile–quantile plots. *The American Statistician*, *70*(1), 74–90.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, *74*(365), 105–121.

Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., & Jin, R. (2019). Softtriple loss: deep metric learning without triplet sampling. (pp. 6450–6458). http://dx.doi.org/10.1109/ICCV.2019.00655,

Reiss, R. D. (2012). *Approximate distributions of order statistics: With applications to nonparametric statistics*. Springer Science & Business Media.

Ren, Y., Zhu, J., Li, J., & Luo, Y. (2016). Conditional generative moment-matching networks. In *Advances in neural information processing systems* (pp. 2928–2936).

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Samaria, F. S., & Harter, A. C. (1994). Parameterisation of a stochastic model for human face identification. In *Proceedings of 1994 IEEE workshop on applications of computer vision* (pp. 138–142). IEEE.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, *100*(5), 401–409.

Saul, L. K., & Roweis, S. T. (2003). Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, *4*(Jun), 119–155.

Schölkopf, B. (2001). The kernel trick for distances. *Advances in Neural Information Processing Systems*, 301–307.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: a unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823). http://dx.doi.org/10.1109/CVPR.2015.7298682.

Scott, D. W. (2015). *Multivariate density estimation: Theory, practice, and visualization*. John Wiley & Sons.

Serfling, R. (2004). Nonparametric multivariate descriptive measures based on spatial quantiles. *Journal of Statistical Planning and Inference*, *123*(2), 259–278.

Sikaroudi, M., Ghojogh, B., Karray, F., Crowley, M., & Tizhoosh, H. R. (2020). Batch-incremental triplet sampling for training triplet networks using Bayesian updating theorem. In *Proceedings of the IEEE international conference on pattern recognition*. IEEE.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Van Der Maaten, L. (2009). Learning a parametric embedding by preserving local structure. In *Artificial intelligence and statistics* (pp. 384–391).

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of Statistics*, 642–656.