

Highlights

Quantile-Quantile Embedding for Distribution Transformation and Manifold Embedding with the Ability to Choose the Embedding Distribution

Benyamin Ghojogh, Fakhri Karray, Mark Crowley

- Proposing a method for distribution transformation and manifold embedding where the user can choose the distribution of embedding.
- Experiments for transforming the distribution of data, by QQE, to any desired distribution while the local distances of data points are preserved.
- Experiments on manifold embedding with QQE where the user can choose the embedding distribution.
- Experiments for changing the distribution of other manifold learning methods, such as PCA, t-SNE, and deep metric learning, to any desired distribution.
- Experiments, for analyzing the effectiveness of the proposed QQE, on both synthetic and real datasets.

Quantile-Quantile Embedding for Distribution Transformation and Manifold Embedding with the Ability to Choose the Embedding Distribution ^{★,★★}

Benyamin Ghoghgh^{a,*}, Fakhri Karray^b and Mark Crowley^a

^aMachine Learning Laboratory, Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

^bCentre 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)
quantilequantile 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 different 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 manners. QQE can also transform distribution to either the exact reference distribution or shape of the reference distribution. [We show that one can also use QQE for better discrimination of classes.](#) Our experiments on different synthetic and image datasets show the effectiveness of the proposed embedding method.

1. Introduction

Data science and machine learning deal with processing data for the sake of data representation, class discrimination, data prediction, and extraction of information from data. Every dataset consists of several data instances. These data instances have a specific probability distribution. The distribution of data may be a standard distribution, e.g. Gaussian distribution, or any strange distribution. Working with data in data science and machine learning may require some pre-processing because the distribution of data may not be suitable for the used data science algorithm or the distribution may not be helpful for discrimination of classes and representation of data. Therefore, in these cases, we need to transform the distribution of data as a pre-processing step to prepare data for other data science and machine learning algorithms.

We can transform data so the distribution of data instances becomes a desired distribution. Note that the relation and local distances of data instances are important because they carry the information of data. Hence, it is important that the data transformation should not significantly modify the relative local distances of nearby data points (Saul and Roweis, 2003). For this distribution transformation, one can try to make all moments of data equal to the moments of the desired distribution (Gretton, Borgwardt, Rasch, Schölkopf and Smola, 2007, 2012). However, because of the huge number of moments, it can be computationally expensive. Furthermore, moments of non-standard distributions can be hard to compute in some cases. Another problem with matching all moments is that it results in transformation to the exact desired distribution but not the “shape” of the desired distribution. One may just want to transform the shape of distribution to the desired one and not to the exact distribution. [Note that transformation of data to the shape of another distribution means that the shape of Probability Density Function \(PDF\) of data becomes similar to the desired PDF regardless of the mean and scale of distribution.](#) Hence, a method for distribution transformation is required which can be used for any desired distribution, either standard or non-standard distributions. [The desired distribution may also be the distribution of some other data instances, named an empirical reference sample.](#) The method for distribution transformation should support a desired distribution either as a theoretical PDF/Cumulative Distribution Function (CDF) or as an empirical reference sample.

 bghoghgh@uwaterloo.ca (B. Ghoghgh); karray@uwaterloo.ca (F. Karray); mcrowley@uwaterloo.ca (M. Crowley)
ORCID(s):

In addition to the necessity of a method for distribution transformation, the need to have the choice of distribution of embedded data is sensed in the field of manifold learning and dimensionality reduction. In other words, one may be interested 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. Some dimensionality reduction methods take an assumption on the distribution of neighbors of data points. Some 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. Therefore, there is a need for a manifold learning and dimensionality reduction method which gives the user the freedom to choose the embedding distribution. Choosing the embedding distribution can be either supervised or unsupervised in which the embedding distribution of entire data or every class is chosen by the user, respectively.

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 choice of embedding distribution. The features and advantages of QQE are summarized in the following:

1. Distribution transformation to a desired distribution either as a PDF/CDF or an empirical reference distribution given by user. Also, either the whole data or every class of data can be transformed in unsupervised and supervised manners, respectively.
2. Manifold embedding of high dimensional data into a lower dimensional embedding space with the choice of embedding distribution by the user. Again, the embedding distribution of either the whole data or every class can be determined in unsupervised and supervised manners, respectively. 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

As distribution transformation transforms the distribution of data to another distribution, it is related to computing the difference of two distributions. There are various methods in statistics for computation of difference of distributions. One of the most well-known methods is the Kullback-Leibler (KL) divergence (Kullback and 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 and Bengio, 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 and 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 and Zemel, 2015; Ren, Zhu, Li and Luo, 2016). Another measure for measuring the relation of two random variables is Hilbert-Schmidt Independence Criterion (HSIC) (Gretton, Bousquet, Smola and 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 and Jahromi, 2011) and supervised guided LLE (Alipanahi and Ghodsi, 2011). Note that the formulas of the three introduced methods for measuring the difference of distributions will be provided in Section 5.2. 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 CDF (Parzen, 1979; Hyndman and Fan, 1996). If we plot the quantile function, we will have the quantile plot (Galton, Foxwell, Martin, Walker, Marshall, Longstaff and Körösi, 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 and Hofmann, 2016). Using the qq-plot, statisticians can visually test whether the two distributions are equal and if not, how much different they are (Oldford, 2016; Loy et al., 2016). There also exist multivariate versions of qq-plot such as fuzzy qq-plot (Easton and 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

As was mentioned in Section 1, the existing manifold learning and dimensionality reduction methods either force an embedding distribution or do not care about it. Some methods take an assumption on the distribution of neighbors of data points. For example, Stochastic Neighbor Embedding (SNE) and t-SNE take Gaussian distribution (Hinton and Roweis, 2003) and Cauchy (Maaten and Hinton, 2008) (or Student-t (Van Der Maaten, 2009)) distribution for the neighborhood of points, respectively. These methods make some strong assumptions on 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 on the embedding distribution and yet do not give any choice of embedding distribution to the user. Some examples are PCA (Ghojogh and Crowley, 2019), Multi-dimensional Scaling (MDS) (Cox and Cox, 2008), Sammon mapping (Sammon, 1969), FDA (Ghojogh, Karray and Crowley, 2019), Isomap (Tenenbaum, De Silva and Langford, 2000), LLE (Roweis and Saul, 2000; Saul and Roweis, 2003), and deep manifold learning (He, Zhang, Ren and Sun, 2016; Schroff, Kalenichenko and Philbin, 2015). Note that some of these methods make assumptions but not as a distribution for the embedding. For example, FDA assumes 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 (Parzen, 1979; Hyndman and Fan, 1996):

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

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

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

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 i -th position, denoted by p_i , as:

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

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 $\mathbf{u}_i \in \mathbb{R}^d$ of $\mathbf{x}_i \in \mathbb{R}^d$ with respect to the sample $\{\mathbf{x}_j\}_{j=1}^n$ is defined as (Möttönen and Oja, 1995; Marden, 2004; Serfling, 2004; Dhar, Chakraborty and Chaudhuri, 2014):

$$\mathbf{u}_i := \frac{1}{n} \sum_{j=1, j \neq i}^n \frac{\mathbf{x}_i - \mathbf{x}_j}{\|\mathbf{x}_i - \mathbf{x}_j\|_2}, \quad (4)$$

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 $\mathbf{u} \in \mathbb{R}^d$ is defined as:

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

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

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 and 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 and McCulloch, 1990). There exist different methods for a promising generalization. One of these methods is *fuzzy qq-plot* (Easton and 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., $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$. We want to transform its distribution as $\mathbf{x}_i \mapsto \mathbf{y}_i$, $\forall i \in \{1, \dots, n\}$. We draw a sample $\{\mathbf{y}_i \in \mathbb{R}^d\}_{i=1}^n$ of size n from the desired (reference) distribution. Note that in case we already have a reference sample $\{\mathbf{y}_i \in \mathbb{R}^d\}_{i=1}^m$ 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 $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$ (Easton and McCulloch, 1990):

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

where $\mathbf{A} \in \mathbb{R}^{d \times d}$ and $\mathbf{b} \in \mathbb{R}^d$ 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.

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

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

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

According to the 1's in the obtained Ψ , we have σ . Then given σ , we solve:

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

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

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

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

Note that it is better to set the initial rotation matrix to the identity matrix, i.e. $\mathbf{A}^{(0)} = \mathbf{I}$, for not having much rotation in assignment. In this way, only few iterations suffice to solve the matching problem. This iterative optimization gives us the matching σ and the samples $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$ are matched. Then, we have d qq-plots, one for every dimension. These qq-plots are named fuzzy qq-plots (Easton and McCulloch, 1990). Considering the spatial ranks, the quantiles are (Dhar et al., 2014):

$$\mathcal{Q}_X(\mathbf{u}_i) = \mathbf{x}_i, \quad \forall i \in \{1, \dots, n\}, \quad (10)$$

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

4. Quantile-Quantile Embedding

In QQE, we want to transform data instances $\{\mathbf{x}_i^0\}_{i=1}^n$ to another dataset $\{\mathbf{x}_i\}_{i=1}^n$ whose distribution is the desired distribution. Now, we provide our definition for distribution transformation:

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

Distribution transformation can be performed in two approaches. In the first approach, (i) the distribution of data is transformed to the “exact” reference distribution, (ii) while in the second approach, only the “shape” of the reference distribution is considered to transform to. In the following Subsections 4.1 and 4.2, we detail the two approaches, respectively. Then, we introduce manifold embedding using QQE in subsection 4.3. Finally, Subsection 4.4 explains the 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 $\{\mathbf{x}_i^0\}_{i=1}^n$ and transform it to $\{\mathbf{x}_i\}_{i=1}^n$ whose distribution is desired to be the same as the distribution of a reference sample $\{\mathbf{y}_{\sigma(i)}\}_{i=1}^n$ or a reference distribution. For this, we consider the fuzzy qq-plot of $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_{\sigma(i)}\}_{i=1}^n$. 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 $\mathcal{Q}_l(\mathbf{u}_i) \in \mathbb{R}$ denote the l -th dimension of $\mathbb{R}^d \ni \mathcal{Q}(\mathbf{u}_i) = [\mathcal{Q}_1(\mathbf{u}_i), \dots, \mathcal{Q}_d(\mathbf{u}_i)]^\top$ which is used for the i -th data point in the l -th qq-plot. Consider $\mathcal{Q}_l(\mathbf{u}_i)$ for the matched data and the reference sample, denoted by $\mathcal{Q}_{X,l}(\mathbf{u}_i)$ and $\mathcal{Q}_{Y,l}(\mathbf{u}_i)$, respectively. In order to have the line in the qq-plot, we should minimize $\sum_{i=1}^n \sum_{l=1}^d (\mathcal{Q}_{X,l}(\mathbf{u}_i) - \mathcal{Q}_{Y,l}(\mathbf{u}_i))^2$. 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 l -th dimension of $\mathbf{x}_i = [x_{i,1}, \dots, x_{i,d}]^\top$ and $\mathbf{y}_{\sigma(i)} = [y_{\sigma(i),1}, \dots, y_{\sigma(i),d}]^\top$, respectively. In vector form, the cost function is restated as:

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

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 and 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 $\{\mathbf{x}_i\}_{i=1}^n$. Let \mathcal{N}_i denote the set containing the indices of the k neighbors of \mathbf{x}_i . The cost to be minimized is:

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

where $d_x(i, j) := \|\mathbf{x}_i - \mathbf{x}_j\|_2$, $d_x^0(i, j) := \|\mathbf{x}_i^0 - \mathbf{x}_j^0\|_2$, 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 (Sammon, 1969; Lee and Verleysen, 2007). We use this cost as a regularization term in our optimization. Therefore, our optimization problem is:

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

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

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

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

Proof. Proof in Appendix A. □

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

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

Proof. Proof in Appendix B. □

We use the quasi-Newton's method (Nocedal and 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 and Verleysen, 2007):

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

$\forall i \in \{1, \dots, n\}, \forall l \in \{1, \dots, d\}$, where v is the index of iteration, $\eta > 0$ is the learning rate, and $|\cdot|$ 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 $\{\mathbf{x}_i^0\}_{i=1}^n$ and transform it to $\{\mathbf{x}_i\}_{i=1}^n$ whose shape of distribution is desired to be similar to the shape of distribution of a reference sample $\{\mathbf{y}_{\sigma(i)}\}_{i=1}^n$. 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 $\mathbb{R}^n \ni \check{\mathbf{Q}}_{Y,l} := [\mathbf{Q}_{Y,l}(\mathbf{u}_1), \dots, \mathbf{Q}_{Y,l}(\mathbf{u}_n)]^\top$, let $\mathbb{R}^{n \times 2} \ni \mathbf{\Gamma}_l := [\mathbf{1}_{n \times 1}, \check{\mathbf{Q}}_{Y,l}]$. Fitting a line to the qq-plot of the l -th dimension is the following least squares problem:

$$\underset{\beta_l}{\text{minimize}} \quad \frac{1}{2} \left\| \mathbf{Q}_X(\mathbf{u}_i) - \mathbf{\Gamma}_l \beta_l \right\|_2^2 \stackrel{(10)}{=} \frac{1}{2} \left\| \mathbf{x}_l - \mathbf{\Gamma}_l \beta_l \right\|_2^2, \quad (18)$$

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

$$\mathbb{R}^2 \ni \beta_l = (\mathbf{\Gamma}_l^\top \mathbf{\Gamma}_l)^{-1} \mathbf{\Gamma}_l^\top \mathbf{x}_l, \quad (19)$$

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

$$\mathbb{R}^n \ni \mu_l := \mathbf{\Gamma}_l \beta_l = [\mu_{\sigma(1),l}, \dots, \mu_{\sigma(n),l}]^\top, \quad (20)$$

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

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

Similar to Proposition 1, the gradient is:

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

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

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 and Crowley, 2019) (or classical MDS (Cox and Cox, 2008)), FDA (Ghojogh et al., 2019), Isomap (Tenenbaum et al., 2000), LLE (Roweis and 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 an *unsupervised* manner, 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 for distribution transformation transforms the entire data to have the desired distribution. The unsupervised QQE for manifold embedding initializes the embedding data in 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 the different classes. The supervised QQE for distribution transformation transforms the distribution of every class to a desired distribution. The supervised QQE for manifold learning initializes the embedding data in 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 no matter 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 repository¹. 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 and Karp, 1972) and the optimization steps, respectively. Improvement of time complexity of QQE is a possible future direction discussed in Section 6.

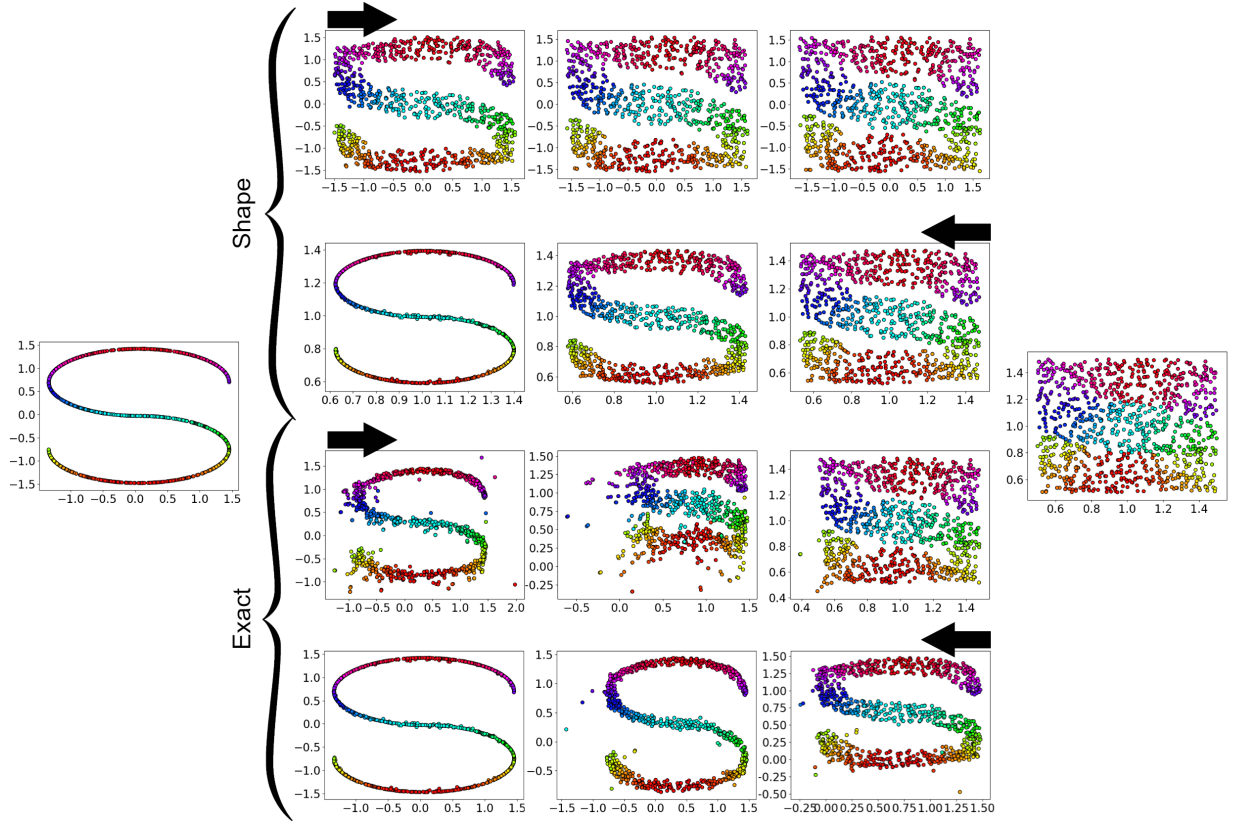


Figure 1: Distribution transformation of S-shape and uniform data to each other. The first and second pair of rows correspond to transformation of shape and exact distributions, respectively. The arrows show the direction of gradual changes.

¹<https://github.com/bgghojogh/Quantile-Quantile-Embedding>

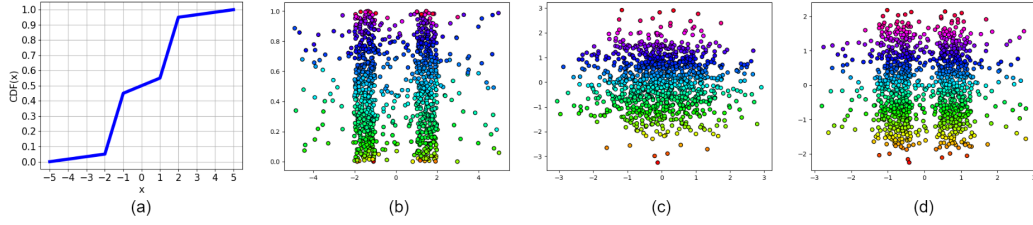


Figure 2: Distribution transformation using (a) CDF of reference distribution: (b) the reference data, (c) Gaussian data, and (d) transformed data.

Table 1

The timing for experiments in this paper. The reported timings are addition of timings for matching and fuzzy qq-plot iterations. In QQE manifold embedding, the time of 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. 3		
Time	206.56	253.49	364.08	278.76	51.57	1064.85		
Experiment	Fig. 4 (PCA, U)	Fig. 4 (PCA, S)	Fig. 4 (FDA, U)	Fig. 4 (FDA, S)	Fig. 4 (Isomap, U)	Fig. 4 (Isomap, S)	Fig. 4 (LLE, U)	Fig. 4 (LLE, S)
Time	206.81	450.87	262.82	269.86	187.53	283.65	328.74	276.10
Experiment	Fig. 4 (t-SNE, U)	Fig. 4 (t-SNE, S)	Fig. 4 (S, Exact)					
Time	289.68	289.62	365.24					
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	2787.25	2527.27	6803.98	2660.45	6654.53	14490.06	4885.59	649.08
Experiment	Fig. 5 (t-SNE, U)	Fig. 5 (t-SNE, S)	Fig. 5 (ResNet, U)	Fig. 5 (ResNet, S)	Fig. 5 (Siamese, U)	Fig. 5 (Siamese, S)	Fig. 5 (S, Exact)	
Time	5373.26	2644.21	6075.75	2332.44	6281.83	30564.37	4893.62	
Experiment	Fig. 7 (synthetic)	Fig. 7 (face)						
Time	365.95	4814.96						

5.1. Discussion on Impact of Hyperparameters

For all experiments in this article, we set $\lambda = 0.1$, $\eta = 0.01$, and $k = 10$. QQE is not yet applicable on out-of-sample data (see Section 6) so these parameters cannot be determined by validation; however, here, we briefly discuss the impact of these hyperparameters. The learning rate η should be set small enough to have progress in optimization without oscillating behaviour. We empirically found $\eta = 0.01$ to be good for different datasets. The larger number of neighbors k results in slower pacing of optimization because of Eqs. (15) and (16). Very small k , however, does not capture the local patterns of data (Saul and Roweis, 2003). The value $k = 10$ is fairly proper. 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. The value $\lambda = 0.1$ was empirically found to be proper for different datasets.

5.2. 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 different quantitative measurements for validating distribution transformation theoretically. Table 2 reports the quantitative measurements for all experiments in this paper. We used three different measures which were initially introduced in Section 2.1. Here, we briefly introduce the formulation of these three measures. These measures are used to show the improvement of change of distributions to the desired distributions using QQE algorithm.

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

$$\text{KL}(\mathcal{P} \parallel \mathcal{Q}) := \sum_{i=1}^n \mathcal{P}(\mathbf{x}_i) \log \left(\frac{\mathcal{P}(\mathbf{x}_i)}{\mathcal{Q}(\mathbf{y}_i)} \right), \quad (23)$$

for the difference of distributions \mathcal{P} and \mathcal{Q} . We estimate $\mathcal{P}(\mathbf{x}_i)$ and $\mathcal{Q}(\mathbf{y}_i)$ using kernel density estimation with Gaussian

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. 3			
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. 4 (PCA, U)	Fig. 4 (PCA, S)	Fig. 4 (FDA, U)	Fig. 4 (FDA, S)	Fig. 4 (Isomap, U)	Fig. 4 (Isomap, S)	Fig. 4 (LLE, U)	Fig. 4 (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. 4 (t-SNE, U)	Fig. 4 (t-SNE, S)	Fig. 4 (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. 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	2.66E-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. 5 (t-SNE, U)	Fig. 5 (t-SNE, S)	Fig. 5 (ResNet, U)	Fig. 5 (ResNet, S)	Fig. 5 (Siamese, U)	Fig. 5 (Siamese, S)	Fig. 5 (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	1.23E0 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. 7 (synthetic)	Fig. 7 (face)							
KL-divergence	2.00E-2 3.13E-2	1.27E-3 1.11E-3							
MMD ²	8.54E-1 3.02E-1	1.23E-2 1.23E-2							
HSIC	4.33E-2 4.80E-2	6.02E-3 6.02E-3							

kernels and the Scott's rule (Scott, 2015). Note that $KL \geq 0$ where $KL = 0$ means the two distributions are equivalent. After applying QQE for distribution transformation or manifold embedding, it is mostly expected to have smaller KL-divergence between the sample $\{\mathbf{x}_i\}_{i=1}^n$ and the reference sample $\{\mathbf{y}_i\}_{i=1}^n$. As KL-divergence does not have any upperbound, the amount of reduction of KL-divergence is not important but the reduction itself is mostly expected.

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 (Hofmann et al., 2008). Let $\phi(\mathbf{x})$ be the pulling function from the input to the feature space and $k(\mathbf{x}_i, \mathbf{x}_j) := \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$ be the kernel function (Hofmann et al., 2008). It is defined as:

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

where $\|\cdot\|$ denotes a norm in the feature/Hilbert space. Note that $\text{MMD} \geq 0$ where $\text{MMD} = 0$ 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 $\{\mathbf{x}_i\}_{i=1}^n$ and the reference sample $\{\mathbf{y}_i\}_{i=1}^n$. As MMD does not have any upperbound, 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 $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ 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 $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$ is (Gretton et al., 2005):

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

where $\text{tr}(\cdot)$ denotes the trace of matrix and \mathbf{K}_x and \mathbf{K}_y are kernels over samples $\{\mathbf{x}_i\}_{i=1}^n$ and $\{\mathbf{y}_i\}_{i=1}^n$, respectively (Hofmann et al., 2008). $\mathbb{R}^{n \times n} \ni \mathbf{H} := \mathbf{I} - (1/n)\mathbf{1}\mathbf{1}^\top$ is the centering matrix where \mathbf{I} and $\mathbf{1}$ denote the identity matrix and the vector of ones, respectively. Note that $\text{HSIC} \geq 0$ where $\text{HSIC} = 0$ 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 $\{\mathbf{x}_i\}_{i=1}^n$ and the reference sample $\{\mathbf{y}_i\}_{i=1}^n$. 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.

5.3. 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.



Figure 3: Distribution transformation of facial images (Samaria and Harter, 1994; Cambridge) without eyeglasses to the shape of images with eyeglasses. The arrow shows the direction of gradual changes.

5.3.1. 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 uniform 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 timing of 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 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 transformation to exact distribution mostly have smaller KL-divergence and MMD and larger HSIC compared to transformation to the shape of reference distribution. This is because exact transformation matches all moments while some moments are not matched in shape transformation.

5.3.2. Given Reference Sample

We can have a reference sample which we want to transform the distribution of data to its distribution. 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 change of moments; thus, some points jump at initial iterations and then converge gradually. In Section 6, we report a future work to make QQE more robust to these jumps. The timing of this experiment is reported in Table 1. The KL-divergence, MMD, and HSIC of distribution transformation of uniform data to either the shape of 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.

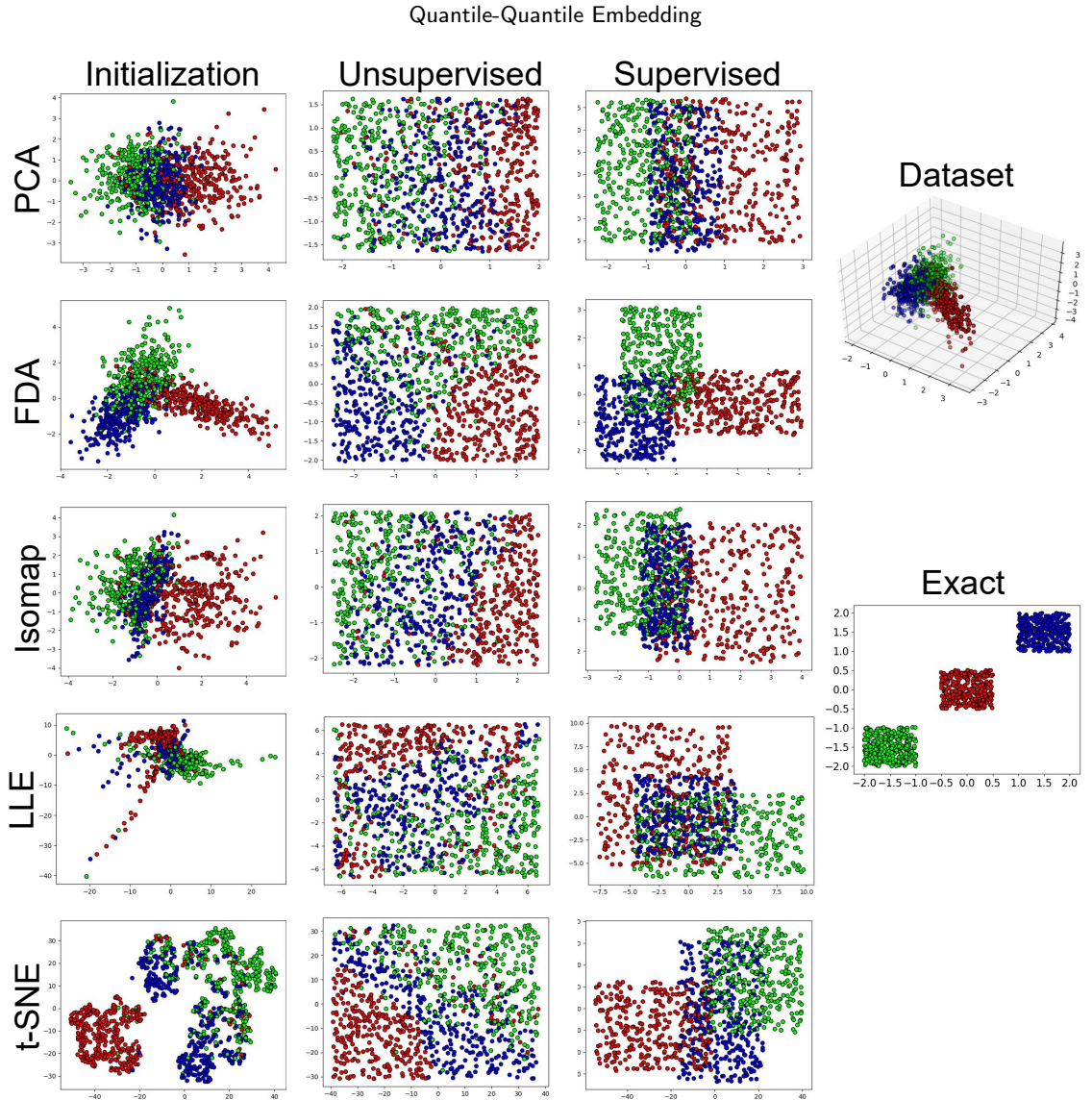


Figure 4: Unsupervised and supervised exact manifold embedding of the synthetic data with different initializations. Transformation to exact reference distribution is also shown. The initialization of LLE is scaled by constant to be in range of other embeddings.

5.3.3. 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, Karay and 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 dimension 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 timing of 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.](#)

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 (Samaria and Harter, 1994; Cambridge) into two sets of with and without eyeglasses. The set with eyeglasses was

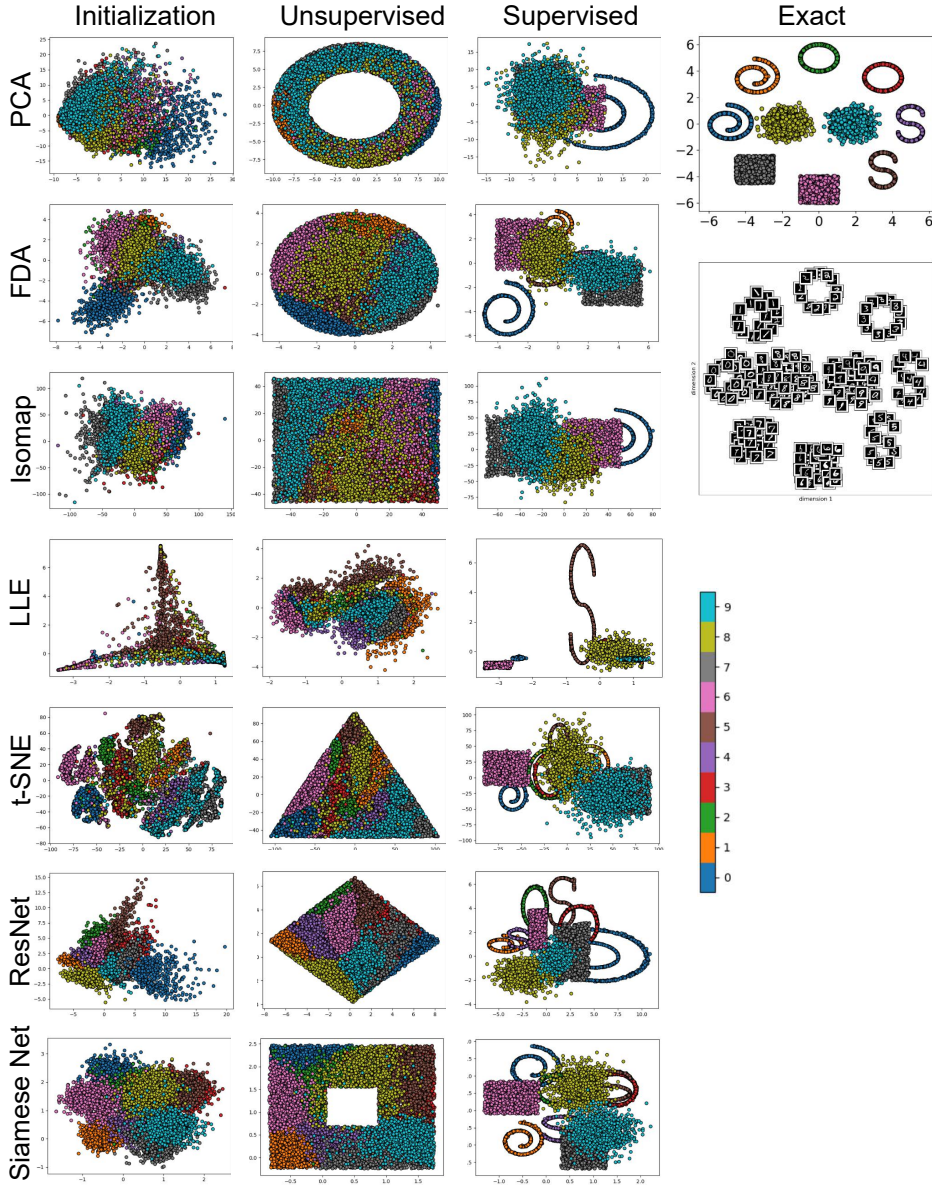


Figure 5: Unsupervised and supervised exact manifold embedding of the image data with different initializations. Transformation to exact reference distribution is also shown. The initialization of LLE is scaled by constant to be in range of other embeddings.

taken as the reference sample and we transformed the set without glasses to have the shape of reference distribution. Figure 3 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 timing of 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.

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. 4. Different dimensionality reduction methods, including PCA (Ghojogh and Crowley, 2019), FDA (Ghojogh et al., 2019), Isomap (Tenenbaum et al., 2000), LLE (Roweis and Saul, 2000), and t-SNE (Van Der Maaten, 2009),

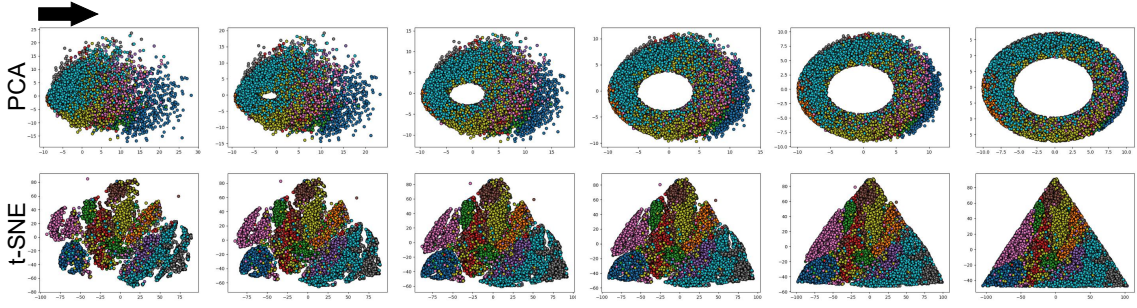


Figure 6: Some iterations of unsupervised manifold embedding initialized by PCA and t-SNE. The arrow shows the direction of gradual changes.

were used for initialization (see the first column in Fig. 4). There are multiple experiments shown in Fig. 4 which we explain in the following:

- For *unsupervised* experiment, we used uniform distribution as reference and transformed the entire embedded data in unsupervised manner. As the second column in Fig. 4 shows, the embeddings of the entire dataset have changed to have the *shape* of uniform distribution but the order and adjacency of classes/points differ according to the initialization methods.
- The results of *supervised* experiments are shown in the third column in Fig. 4. The desired reference distribution for every class was uniform distribution and we desired the *shape* of 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. 4 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 timing of 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 and 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 and Crowley, 2019), FDA (Ghojogh et al., 2019), Isomap (Tenenbaum et al., 2000), LLE (Roweis and 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. 5.

Figure 5 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. 5, the shape of entire embedding has changed to the desired while the local distances are preserved as much as possible. Figure 6 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. 5 depicts the supervised transformation to shapes of distributions.

- The fourth column in Fig. 5 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. 5 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 timing of 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.

5.7. 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. 7. 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 and 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. 7, were obtained using the Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy and 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 timing of both 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 3 for both synthetic and image data. Following the literature (Qian, Shang, Sun, Hu, Li and Jin, 2019; Nguyen and De Baets, 2020; Sikaroudi, Ghogh, Karray, Crowley and Tizhoosh, 2020), we used the Recall@k measure for supervised evaluation of embedding in terms of discrimination of classes. As this table show, QQE has improved the separation of classes by its steps. This improvement in separation of classes can also be seen in Fig. 7.

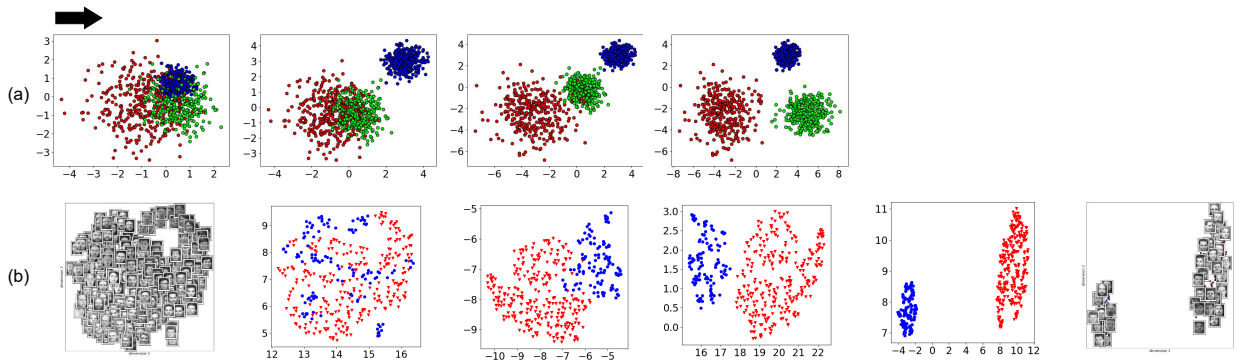


Figure 7: Separation and discrimination of classes in synthetic and image data. The arrow shows the direction of gradual changes.

5.8. Evaluating QQE for Histopathology data

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 shape of it. Both unsupervised and supervised versions of

Table 3

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 on the steps of experiments shown in Fig. 7.

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

this method were also proposed. The proposed method was based on quantile-quantile plot which is usually used in visual statistical tests.

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.

A. Proof of Proposition 1

Consider the first part of the cost function:

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

Consider the second part of the cost function:

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

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:

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

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

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

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

Considering both parts of the cost function, the gradient is as in the proposition. Q.E.D.

B. Proof of Proposition 2

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

$$\begin{aligned} \frac{\partial^2 \mathcal{L}}{\partial x_{i,l}^2} &= 1 + \frac{\lambda}{a} \sum_{j \in \mathcal{N}_i} \frac{\partial}{\partial x_{i,l}} \left(\frac{d_x(i, j) - d_x^0(i, j)}{d_x(i, j) d_x^0(i, j)} (x_{i,l} - x_{j,l}) \right) \\ &= \frac{\partial}{\partial x_{i,l}} \left(\frac{d_x(i, j) - d_x^0(i, j)}{d_x(i, j) d_x^0(i, j)} (x_{i,l} - x_{j,l}) \right) \\ &= (x_{i,l} - x_{j,l}) \frac{\partial}{\partial x_{i,l}} \left(\frac{d_x(i, j) - d_x^0(i, j)}{d_x(i, j) d_x^0(i, j)} \right) + \underbrace{\frac{d_x(i, j) - d_x^0(i, j)}{d_x(i, j) d_x^0(i, j)} \frac{\partial}{\partial x_{i,l}} (x_{i,l} - x_{j,l})}_{=1} \\ &= \frac{\partial}{\partial x_{i,l}} \left(\frac{d_x(i, j) - d_x^0(i, j)}{d_x(i, j) d_x^0(i, j)} \right) = \frac{1}{d_x^0(i, j)} \frac{\partial}{\partial x_{i,l}} \left(1 - \frac{d_x^0(i, j)}{d_x(i, j)} \right) \\ &= \frac{1}{d_x^0(i, j)} \underbrace{\frac{\partial}{\partial x_{i,l}} (1)}_{=0} - \underbrace{\frac{d_x^0(i, j)}{d_x^0(i, j)} \frac{\partial}{\partial x_{i,l}} \left(\frac{1}{d_x(i, j)} \right)}_{=1} = \frac{-1}{d_x^2(i, j)} \frac{\partial}{\partial x_{i,l}} (d_x(i, j)) \stackrel{(26)}{=} \frac{-(x_{i,l} - x_{j,l})}{d_x^3(i, j)}. \end{aligned}$$

Putting all parts of derivative together gives the second derivative. Q.E.D.

CRedit authorship contribution statement

Benyamin Ghogh: Conceptualization of this study, Methodology, Software, Writing, Discussion. **Fakhri Kar-ray:** Supervision, Writing, Discussion. **Mark Crowley:** Supervision, Writing, Discussion.

References

- Alipanahi, B., Ghodsi, A., 2011. Guided locally linear embedding. *Pattern Recognition Letters* 32, 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, 1357–1371.
- Cambridge, A.L., . AT&T laboratories Cambridge. <http://cam-orl.co.uk/facedatabase.html>. Accessed: 2020-01-01.
- Chaudhuri, P., 1996. On a geometric notion of quantiles for multivariate data. *Journal of the American Statistical Association* 91, 862–872.
- Cox, M.A., Cox, T.F., 2008. Multidimensional scaling, in: *Handbook of data visualization*. Springer, pp. 315–347.
- Dhar, S.S., Chakraborty, B., Chaudhuri, P., 2014. Comparison of multivariate distributions using quantile–quantile plots and related tests. *Bernoulli* 20, 1484–1506.
- Easton, G.S., McCulloch, R.E., 1990. A multivariate generalization of quantile-quantile plots. *Journal of the American Statistical Association* 85, 376–386.
- Edmonds, J., Karp, R.M., 1972. Theoretical improvements in algorithmic efficiency for network flow problems. *Journal of the ACM (JACM)* 19, 248–264.
- Ferguson, T.S., 1967. *Mathematical statistics: A decision theoretic approach*. Academic press.
- Galton, F., 1874. On a proposed statistical scale. *Nature* 9, 342.
- Galton, F., Foxwell, P., Martin, J.B., Walker, F., Marshall, P., Longstaff, G., Körösi, H., 1885. The application of a graphic method to fallible measures [with discussion]. *Journal of the Statistical Society of London* , 262–271.
- Ghogh, 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 .
- Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y., 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, 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, Springer. pp. 63–77.
- 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.
- 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, 361–365.
- Kuhn, H.W., 1955. The Hungarian method for the assignment problem. Naval research logistics quarterly 2, 83–97.
- Kullback, S., Leibler, R.A., 1951. On information and sufficiency. The annals of mathematical statistics 22, 79–86.
- LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 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, 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, 202–214.
- Maaten, L.V.D., Hinton, G., 2008. Visualizing data using t-SNE. Journal of machine learning research 9, 2579–2605.
- Marden, J.I., 2004. Positions and QQ plots. Statistical Science 19, 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. Journaltitle of Nonparametric Statistics 5, 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, 74–90.
- Parzen, E., 1979. Nonparametric statistical data modeling. Journal of the American statistical association 74, 105–121.
- Qian, Q., Shang, L., Sun, B., Hu, J., Li, H., Jin, R., 2019. Softtriplet loss: Deep metric learning without triplet sampling, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6450–6458.
- 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, 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, IEEE. pp. 138–142.
- Sammon, J.W., 1969. A nonlinear mapping for data structure analysis. IEEE Transactions on computers 100, 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, 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.
- 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, 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 (ICPR), IEEE.
- Tenenbaum, J.B., De Silva, V., Langford, J.C., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 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.