Calibrating Distance Metrics Under Uncertainty

Wenye Li[1,2(✉)][0000−0002−5679−9670] and Fangchen Yu[1][0000−0002−1256−2719]

[1] The Chinese University of Hong Kong, Shenzhen, China
wyli@cuhk.edu.cn, fangchenyu@link.cuhk.edu.cn
[2] Shenzhen Research Institute of Big Data, Shenzhen, China

Abstract. Estimating distance metrics for given data samples is essential in machine learning algorithms with various applications. Accurately determining the metric becomes impossible if there are observation noises or missing values. In this work, we proposed an approach to calibrating distance metrics. Compared with standard practices that primarily reside on data imputation, our proposal makes fewer assumptions about the data. It provides a solid theoretical guarantee in improving the quality of the estimate. We developed a simple, efficient, yet effective computing procedure that scales up to realize the calibration process. The experimental results from a series of empirical evaluations justified the benefits of the proposed approach and demonstrated its high potential in practical applications.

Keywords: Missing Data · Metric Calibration · Alternating Projection.

1 Introduction

In data processing, a distance metric, or a distance matrix, is used to measure the pairwise dis-similarity relationship between data samples. It is crucial and lays a foundation in many supervised and unsupervised learning models, such as the K-means clustering algorithm, the nearest neighbor classifier, support vector machines [18,9,30,35].

Calculating pairwise distance is straightforward if the data samples are clean and fully observed. Unfortunately, with observation noises or missing values, which are natural and common in practice, obtaining a high-quality distance metric becomes a challenging task, and nontrivial challenges arise to learning algorithms based on the distance estimation between data samples.

Significant research attention has been devoted to handling the difficulty brought by missing values. Various imputation techniques were designed as a routine treatment, which has greatly influenced the progress in various disciplines [25,11]. These techniques complete the data by replacing the missing values with substituted ones based on various assumptions. Based on the imputed data, the pairwise distances can be calculated accordingly.

Despite the popularity received by data imputation approaches, nontrivial challenges still exist. When the assumptions made by the imputation techniques are violated, there is no guarantee at all on the quality of the imputed values, needless to say, the impact on subsequent data analysis tasks. Furthermore, with a large portion of missing values, the imputation can be highly demanding or even prohibitive in computation, which becomes another serious concern.

As a remedy, we carried out a series of work in two directions. Firstly and as the main contribution of this paper, our work proposed a metric calibration model that avoids data imputation in Section 3.2. It starts from an approximate metric estimated from incomplete samples or prior knowledge and then calibrates the metric iteratively. The calibrated metric is guaranteed to be better than the initial metric in terms of a shorter Frobenius distance to the accurate unknown metric, except in rare cases, the two metrics are identical. Secondly, our work applied Dykstra's projection algorithm to realize the calibration process and designed a cyclic projection algorithm as a more scalable alternative.

Compared with the popular imputation methods in handling missing data, the calibration approaches seemed to rely less on the assumption of the correlation among data features or the data's intrinsically low dimension/rank. As a result of the less dependency, the approaches reported more robust and reliable results in empirical evaluations. The improvement from the calibration approaches is especially significant when the missing ratio is high, or the noisy level is high, which exhibited their high potential in handling missing and noisy data in practical applications.

The paper is organized as follows. Section 2 introduces the background. Section 3 presents our model and algorithms. Section 4 reports the experimental results, followed by the conclusion in Section 5.

2 Background

2.1 Missing Data and Imputation

Missing observations are everywhere and pose nontrivial challenges to numerous data analysis applications in science and engineering. Developing techniques to process incompletely observed data becomes one of the most critical tasks in statistical sciences [25,11].

A common approach to dealing with missing observations is through data imputation. A missing value may be replaced by a zero value, the feature's mean, median, or the most frequent value among the nearest neighbor samples or all observed samples.

A more rigorous treatment is based on the expectation-maximization (EM) algorithm [6]. The approach assumes the existence of specific latent structures and variables. By alternatively estimating the model parameters and the missing values with the fitted model parameters, the approach generates a maximum likelihood or a maximum a posterior estimate for each missing observation.

Another imputation approach, the low-rank matrix completion approach developed more recently, makes assumptions on the rank of the data matrix to be completed. Efficient algorithms were designed to achieve exact reconstruction with high probability and reported quite successful results, such as in recommender systems [3,19]. In recent work, based on the assumption that two random batches from the same dataset share the same distribution, a measure of optimal transport distances is applied as an optimization objective for missing data imputation, which achieves excellent performances on some practical tasks [27].

Despite the success and the popularity that has been achieved, an inherent challenge exists. All imputation approaches, either explicitly or implicitly, have assumed the low dimensionality or the low-rank structure of the data. However, when the assumption does not hold, all these approaches will lose the performance guarantee on the imputation quality.

2.2 Metric Calibration

Instead of imputations, a matrix calibration approach can be applied to improve a metric obtained from incomplete or noisy data. As an example, let us consider the *metric nearness* model [2]. Denote the set of all $n \times n$ matrices by $\mathcal{M}_n$, which is a closed, convex polyhedral cone. Assume we are given n incomplete samples and an estimate of their distance matrix $D^0 = \left\{d_{ij}^0\right\}_{i,j=1}^n \in \mathcal{M}_n$. The estimate is inaccurate and might violate the triangle inequality property that the true metric possesses. As a remedy, we consider the following model:

$$\min_{D \in \mathcal{M}_n} \left\| D - D^0 \right\|_F^2 \tag{1}$$

s.t.,

$$d_{ij} \geq 0,\ d_{ii} = 0,\ d_{ij} = d_{ji},\ \text{and}\ d_{ij} \leq d_{ik} + d_{kj},$$

for all $1 \leq i, j, k \leq n$.

The model above seeks a new matrix $D = \{d_{ij}\}_{i,j=1}^n$ that *best approximates* the input matrix D^0 in Frobenius norm, from a feasible region of matrices that meet the desired constraints. After calibration, the result will restore the property that the true distance metric should possess.

The calibration approach has an implicit but key benefit [23]. Suppose the feasible region of the distance matrix of interest is appropriately defined. In that case, although the factual matrix is never known to us, the new calibrated matrix can be guaranteed to be nearer to the ground truth than the initial estimate D^0, except in rare cases that they are identical.

The metric nearness model defined in Eq. (1) can be formulated as a quadratic program and solved by modern convex optimization packages [1]. Besides, an elegant *triangle fixing* algorithm [2] was developed, which exploited the inherent structure of the triangle inequalities and improved running efficiency. Besides, we can also consider a stochastic sampling of constraints or Lagrangian formulations to seek an algorithmic solution [31]. Despite the partial success that has been achieved along this line, however, the intrinsic complexity from $O\left(n^3\right)$ inequality constraints to Eq. (1) makes the model hard to scale up, which significantly limits the application of the model.

3 Model

3.1 A Kernel's Trick

Our work resides on a mild assumption that the data samples in the study are isometrically embeddable in a real Hilbert space, or equivalently, the samples can be represented as real vectors. Recall the definition of isometrical embedding.

Definition 1. *Consider a separable metric space $\mathcal{X}$ with a distance function ρ, having the properties that $\rho(x, x') = \rho(x', x) \geq 0$ and $\rho(x, x) = 0$ for all points x and x' in $\mathcal{X}$. $(\mathcal{X}, \rho)$ is said to be isometrically embeddable in a real Hilbert space $\mathcal{H}$ (or embeddable, for short) if there exists a map $\phi : \mathcal{X} \mapsto \mathcal{H}$ such that*

$$\|\phi(x) - \phi(x')\| = \rho(x, x')$$

for all points x and x' in $\mathcal{X}$.

A classical result on isometrical embedding [29,34] states that:

Theorem 2. *Assume $(\mathcal{X}, \rho)$ is embeddable. Then, for each $\gamma > 0$ and $0 < \alpha < 1$,*

$$\sum_{i,j=1}^{n} \exp\left(-\gamma \rho^{2\alpha}(x_i, x_j)\right) \xi_i \xi_j \geq 0$$

holds for every choice of points $x_1, \cdots, x_n$ in $\mathcal{X}$ and real $\xi_1, \cdots, \xi_n$.

For any finite subset $\{x_1, \cdots, x_n\} \subseteq \mathcal{X}$ $(n \geq 2)$, denote by $D^* = \left\{d^*_{ij}\right\}_{i,j=1}^{n}$ with $d^*_{ij} = \rho(x_i, x_j)$ for each i, j and $\exp\left(-\gamma D^*\right) = \left\{\exp\left(-\gamma d^*_{ij}\right)\right\}_{i,j=1}^{n}$. By choosing $\alpha = \frac{1}{2}$, we have, if $(\mathcal{X}, \rho)$ is embeddable, the matrix $\exp\left(-\gamma D^*\right)$ is positive semi-definite and we also say that the matrix D^* is embeddable.

In the machine learning area, the positive definite function $\exp\left(-\gamma \|x - x'\|\right)$ is known as the Laplacian kernel, popularly used in the context of kernel-based algorithms [30]. In our application, the function connects an embeddable metric D^* and a positive semi-definiteness matrix $\exp\left(-\gamma D^*\right)$.

3.2 Direct Calibration

For given samples $\{x_1, \cdots, x_n\}$ in $\mathcal{X}$, let $D^0 = \left\{d^0_{ij}\right\}_{i,j=1}^{n}$ be an input distance matrix between the samples. Assume that, due to observation noise or missing values, the metric D^0 is not accurate. From the relationship between isometrical embedding and positive semi-definiteness in Section 3.1, we naturally investigate the following model to calibrate the matrix D^0 to a better estimate:

$$\min_{D \in \mathcal{M}_n} \left\|D - D^0\right\|_F^2, \tag{2}$$

s.t.

$$\exp(-\gamma D) \succeq 0, \ \ d_{ii} = 0 \ \ (1 \leq i \leq n), \ \text{and} \ d_{ij} \geq 0 \ \ (1 \leq i \neq j \leq n),$$

where $\succeq 0$ denotes the positive semi-definiteness constraint on a matrix.

Solving the optimization problem in Eq. (2) is not straightforward. Here we develop an efficient approximation. Let $\mu = \max\{\rho(x_i, x_j), 1 \leq i, j \leq n\}$ be a normalizing factor, and $\gamma = \frac{\epsilon}{\mu}$ where ϵ is a small positive number[1]. Denote $E^0 = \left\{e_{ij}^0\right\}_{i,j=1}^n = \exp(-\gamma D^0)$, and we reach a known problem in literature [23]:

$$\min_{E \in \mathcal{M}_n} \left\|E - E^0\right\|_F^2 \tag{3}$$

s.t.

$$E \succeq 0, \;\; e_{ii} = 1 \;\; (1 \leq i \leq n), \text{ and } e_{ij} \in [1-\epsilon, 1] \;\;\; (1 \leq i \neq j \leq n).$$

Let $\mathcal{R} = \{X \in \mathcal{M}_n | X \succeq 0, x_{ii} = 1, x_{ij} \in [1-\epsilon, 1] \text{ for all } i, j\}$ be a closed convex subset of $\mathcal{M}_n$. The optimal solution to Eq. (3) is the projection of E^0 onto $\mathcal{R}$, denoted by $E^0_{\mathcal{R}}$. Let $E^* = \exp(-\gamma D^*)$, where D^* is the true but unknown metric. Obviously $E^* \in \mathcal{R}$, and

$$\begin{aligned}\left\|E^* - E^0_{\mathcal{R}}\right\|_F^2 &\leq \left\|E^* - E^0_{\mathcal{R}}\right\|_F^2 - 2\left\langle E^* - E^0_{\mathcal{R}}, E^0 - E^0_{\mathcal{R}}\right\rangle \\ &\leq \left\|\left(E^* - E^0_{\mathcal{R}}\right) - \left(E^0 - E^0_{\mathcal{R}}\right)\right\|_F^2 .\end{aligned} \tag{4}$$

The first "$\leq$" holds due to Kolmogrov's criterion [7], which states that the projection of E^0 onto $\mathcal{R}$ is unique and characterized by:

$$E^0_{\mathcal{R}} \in \mathcal{R} \text{ and } \left\langle E - E^0_{\mathcal{R}}, E^0 - E^0_{\mathcal{R}}\right\rangle \leq 0, \text{ for all } E \in \mathcal{R}.$$

The equality holds if and only if $E^0_{\mathcal{R}} = E^0$, i.e., $E^0 \in \mathcal{R}$.

Eq. (4) gives $\left\|E^* - E^0_{\mathcal{R}}\right\|_F^2 \leq \left\|E^* - E^0\right\|_F^2$, which shows that $E^0_{\mathcal{R}}$ is an improved estimate towards the unknown E^*. Next, let $D^0_{\mathcal{R}}$ be obtained from $E^0_{\mathcal{R}} = \exp(-\gamma D^0_{\mathcal{R}})$. From Taylor-series expansion:

$$e^z = 1 + z + O\left(z^2\right) \approx 1 + z, \text{ for } |z| \ll 1,$$

we have:

$$\left\|E^* - E^0_{\mathcal{R}}\right\|_F^2 = \frac{\epsilon^2}{\mu^2}\left\|D^* - D^0_{\mathcal{R}}\right\|_F^2 + O(\epsilon^4), \tag{5}$$

and

$$\left\|E^* - E^0\right\|_F^2 = \frac{\epsilon^2}{\mu^2}\left\|D^* - D^0\right\|_F^2 + O(\epsilon^4). \tag{6}$$

If $E^0 \notin \mathcal{R}$, we have $\left\|E^* - E^0_{\mathcal{R}}\right\|_F^2 < \left\|E^* - E^0\right\|_F^2$. For a sufficiently small ϵ, we have $\left\|D^* - D^0_{\mathcal{R}}\right\|_F^2 < \left\|D^* - D^0\right\|_F^2$ based on Eqs. (5) and (6). If $E^0 \in \mathcal{R}$, we

[1] We set $\mu = \max\left\{d_{ij}^0\right\}$ and $\epsilon = 0.02$ in the study.

have $E^0_{\mathcal{R}} = E^0$, which implies $D^0_{\mathcal{R}} = D^0$. Considering both cases, we have:

$$\left\| D^* - D^0_{\mathcal{R}} \right\|_F^2 \leq \left\| D^* - D^0 \right\|_F^2 . \tag{7}$$

The equality holds if and only if $\exp(-\lambda D^0) \succeq 0$. The result shows that, except in the special case $D^0_{\mathcal{R}} = D^0$ that happens when $E^0 \in \mathcal{R}$, the calibrated $D^0_{\mathcal{R}}$ is a better estimate than the input D^0 in terms of a smaller Frobenius distance to the true but unknown metric D^*.

3.3 Dykstra's Algorithm

Solving Eq. (3) to find the projection of E^0 onto set $\mathcal{R}$ is well-studied in the optimization community. Several algorithms are available with quite good performances [28]. Similarly to the work of [16,23], we resort to a simple and flexible procedure based on Dykstra's alternating projection algorithm [10], also called *direct calibration* in the sequel.

Equip the closed convex set $\mathcal{M}_n$ with an inner product that induces the Frobenius norm:

$$\langle X, Y \rangle = trace\left(X^T Y \right), \text{ for } X, Y \in \mathcal{M}_n.$$

Define two nonempty, closed and convex subsets of $\mathcal{M}_n$:

$$\mathcal{S} = \{ X \in \mathcal{M}_n | X \succeq 0 \}, \text{ and } \mathcal{T} = \{ X \in \mathcal{M}_n | x_{ii} = 1, x_{ij} \in [1-\epsilon, 1] \text{ for all } i, j \}.$$

Obviously $\mathcal{R} = \mathcal{S} \cap \mathcal{T}$. Directly projecting E^0 onto $\mathcal{R}$ is expensive, while projecting it onto $\mathcal{S}$ and $\mathcal{T}$ respectively is easier. Denote by $\mathrm{P}_{\mathcal{S}}$ the projection onto $\mathcal{S}$, and $\mathrm{P}_{\mathcal{T}}$ the projection onto $\mathcal{T}$. For $\mathrm{P}_{\mathcal{S}}$ and $\mathrm{P}_{\mathcal{T}}$, we have the following two results.

Fact 1 *Let $X \in \mathcal{M}_n$ and $U \Sigma V^T$ be its singular value decomposition with $\Sigma = diag\left(\lambda_1, \cdots, \lambda_n\right)$. The projection of X onto $\mathcal{S}$ is given by: $X_{\mathcal{S}} = \mathrm{P}_{\mathcal{S}}(X) = U \Sigma' V^T$ where $\Sigma' = \mathrm{diag}\left(\lambda'_1, \cdots, \lambda'_n\right)$ and each $\lambda'_i = \max\{\lambda_i, 0\}$.*

Fact 2 *The projection of $X \in \mathcal{M}_n$ onto $\mathcal{T}$ is given by: $X_{\mathcal{T}} = \mathrm{P}_{\mathcal{T}}(X) = \left\{ \left(x_{\mathcal{T}}\right)_{ij} \right\}_{i,j=1}^n$ where $\left(x_{\mathcal{T}}\right)_{ij} = \mathrm{med}\{1-\epsilon, x_{ij}, 1\}$, i.e., the median of the three numbers.*

Dykstra's projection algorithm can be applied to find the minimizer to Eq. (3). Starting from E^0, it generates a sequence of iterates $\left\{ E^t_{\mathcal{S}}, E^t_{\mathcal{T}} \right\}$ and increments $\left\{ I^t_{\mathcal{S}}, I^t_{\mathcal{T}} \right\}$, for $t = 1, 2, \cdots$, by:

$$E^t_{\mathcal{S}} = \mathrm{P}_{\mathcal{S}}\left(E^{t-1}_{\mathcal{T}} - I^{t-1}_{\mathcal{S}} \right) \tag{8}$$

$$I^t_{\mathcal{S}} = E^t_{\mathcal{S}} - \left(E^{t-1}_{\mathcal{T}} - I^{t-1}_{\mathcal{S}} \right) \tag{9}$$

$$E^t_{\mathcal{T}} = \mathrm{P}_{\mathcal{T}}\left(E^t_{\mathcal{S}} - I^{t-1}_{\mathcal{T}} \right) \tag{10}$$

$$I^t_{\mathcal{T}} = E^t_{\mathcal{T}} - \left(E^t_{\mathcal{S}} - I^{t-1}_{\mathcal{T}} \right) \tag{11}$$

where $E^0_{\mathcal{T}} = E^0$, $I^0_{\mathcal{S}} = \mathbf{0}$, $I^0_{\mathcal{T}} = \mathbf{0}$ and $\mathbf{0}$ is an all-zero matrix of proper size. The sequences $\left\{E^t_{\mathcal{S}}\right\}$ and $\left\{E^t_{\mathcal{T}}\right\}$ converge to the optimal solution $E^0_{\mathcal{R}}$ as $t \to \infty$.

3.4 Cyclic Calibration

Based on the proposed calibration model and the Dykstra's alternating projection algorithm presented in Sections 3.2 and 3.3 respectively, a more scalable calibration algorithm, called *cyclic calibration* [24] in the sequel, can be designed based on the following result.

Fact 3 *Let $\mathcal{R}$ be a closed convex subset of $\mathcal{M}_n$ and $E^* \in \mathcal{R}$. Let $\mathcal{C}$ be a closed convex superset of $\mathcal{R}$ and $\mathcal{C} \subseteq \mathcal{M}_n$. For any $E^0 \in \mathcal{M}_n$, we have $\left\|E^* - E^0_{\mathcal{C}}\right\|^2_F \leq \left\|E^* - E^0\right\|^2_F$. The equality holds if and only if $E^0_{\mathcal{C}} = E^0$, i.e., $E^0 \in \mathcal{C}$.*

This result can be obtained similarly to Eq. (4). It states that the projection of E^0 onto $\mathcal{C}$ provides an improved estimate towards E^*. Based on the observation, we can design a domain decomposition algorithm that avoids factorizing the full $n \times n$ matrix. Let $\mathcal{C}_1, \cdots, \mathcal{C}_r$ be r closed convex sets that satisfy $\mathcal{R} \subseteq \bigcap_{k=1}^r \mathcal{C}_k$ and $\bigcup_{k=1}^r \mathcal{C}_k \subseteq \mathcal{M}_n$. Starting from $E^0 \in \mathcal{M}_n$, again we apply Dykstra's projection which generates the iterates $\{E^t_k\}$ and the increments $\{I^t_k\}$ cyclically by:

$$E^t_0 = E^{t-1}_r \tag{12}$$

$$E^t_k = \mathrm{P}_{\mathcal{C}_k}\left(E^t_{k-1} - I^{t-1}_k\right) \tag{13}$$

$$I^t_k = E^t_k - \left(E^t_{k-1} - I^{t-1}_k\right) \tag{14}$$

where $k = 1, \cdots, r$ and $t = 1, 2, \cdots$. The initial values are given by $E^0_r = E^0$ and $I^0_k = \mathbf{0}$ $(1 \leq k \leq r)$. The sequences of $\{E^t_k\}$ converges to the projection of E^0 onto $\bigcap_{k=1}^r \mathcal{C}_k$ [10].

Theorem 3. *Let $\mathcal{C}_1, \cdots, \mathcal{C}_r$ be closed and convex subsets of $\mathcal{M}_n$ such that $\mathcal{C} = \bigcap_{k=1}^r \mathcal{C}_k$ is not empty. For any $E^0 \in \mathcal{M}_n$ and any $k = 1, \cdots, r$, the sequence $\{E^t_k\}$ converges strongly to $E^0_{\mathcal{C}} = \mathrm{P}_{\mathcal{C}}\left(E^0\right)$, i.e., $\left\|E^t_k - E^0_{\mathcal{C}}\right\|^2_F \to 0$ as $t \to \infty$.*

To realize the cyclic calibration approach, we define the r supersets $\mathcal{C}_1, \cdots, \mathcal{C}_r$ of $\mathcal{R}$ as follows. Denote r nonempty index sets by $\mathcal{I}_1, \cdots, \mathcal{I}_r$, which satisfies $\bigcup_{k=1}^r \mathcal{I}_k = \{1, \cdots, n\}$. For any matrix $A \in \mathcal{M}_n$, denote by A_k the principal submatrix formed by selecting the same rows and columns of A indicated by $\mathcal{I}_k$. Then for each $\mathcal{I}_k$ $(1 \leq k \leq r)$, define

$$\mathcal{S}_k = \{A \in \mathcal{M}_n | A_k \succeq 0\} \text{ , and, } \mathcal{C}_k = \mathcal{S}_k \cap \mathcal{T}.$$

Recall that a matrix is positive semi-definite if and only if all its principal submatrices are positive semi-definite [14,17], and we know that $\mathcal{R} \subseteq \mathcal{C} = \bigcap_{k=1}^r \mathcal{C}_k$. So by projecting E^0 onto each $\mathcal{C}_k$ successively with Dykstra's procedure, we will obtain the projection onto $\mathcal{C}$, which provides an improved estimate towards the unknown E^*, with the following steps:

1. For given r, randomly generate index sets $\mathcal{I}_1, \ldots, \mathcal{I}_r$;
2. Calibrate the matrix by projecting it onto $\mathcal{C}_1, \cdots, \mathcal{C}_r$ cyclically;
3. Repeat steps 1 and 2 until convergence.

Cyclic calibration can be regarded as an extension of the direct calibration presented in Section 3.3. When $r = 1$, the cyclic algorithm reduces exactly to the direct algorithm.

Let $D^0_{\mathcal{C}}$ be obtained from $E^0_{\mathcal{C}}$. Similarly to the result in Eq. (7), we have:

$$\left\| D^* - D^0_{\mathcal{C}} \right\|_F^2 \leq \left\| D^* - D^0 \right\|_F^2 , \tag{15}$$

which shows that $D^0_{\mathcal{C}}$ improves D^0 and gets nearer to the unknown D^*.

3.5 Complexity Analysis

To project an $n \times n$ matrix directly onto the convex set $\mathcal{S}$ via SVD, the complexity is $O\left(n^3\right)$ per iteration [15,5]. With cyclic calibration, we set the cardinality of I_k to $O\left(n/r\right)$ and project an input matrix onto $\mathcal{S}_k$ $(1 \leq k \leq r)$ successively. We need to decompose r principal submatrices in each iteration. The complexity is $O\left(n^3/r^3\right)$ to decompose one submatrix, and $O\left(n^3/r^2\right)$ for r decompositions, which significantly improves the complexity of the direct approach.

For the number of iterations to converge, theoretically, the convergence rate of Dykstra's alternating projection for polyhedral sets is known to be linear [10,12]. Empirically the direct approach converged in around 20 iterations, and the cyclic approach converged in around 40 iterations on a problem with $n = 10,000$ and $r = 10$ in our evaluation.

For memory requirement, if the whole distance matrix is stored in memory, both calibration approaches have a storage complexity of $O\left(n^2\right)$. For the cyclic approach, it is also possible to reduce the storage complexity to $O\left(n^2/r^2\right)$ by only keeping the working principal submatrix in memory, at the cost of swapping-in and swapping-out operations on other matrix elements from time to time.

4 Evaluation

4.1 Settings

We carried out empirical studies to evaluate the proposed model and calibration algorithms, specifically with the objectives of investigating their effectiveness in:

- reducing the noise of distance matrices;
- computing distance metrics from incomplete data;
- performances in classification applications;
- running speed and scalability.

We used five benchmark datasets that are publicly available. These datasets cover a reasonably wide range of application domains, including:

Table 1. Relative Squared Deviations on Calibration of Noisy Distance Metrics. Each item has two values, corresponding to $\zeta = 0.1$ and 0.5 respectively: the smaller RSD value, the better performance. Direct calibration reported the best calibration quality on almost all experiments.

Dataset	TRIFIX	DIRECT	CYCLIC
MNIST	.976/.362	**.137/.034**	.164/.039
CIFAR10	.904/.369	**.146/.031**	.181/.037
PROTEIN	.992/.358	**.178**/.031	.222/**.025**
RCV1	.999/.356	**.184/.023**	.233/.023
SENSEIT	.778/.351	**.103/.032**	.123/.045

- MNIST: images of handwritten digits with 28 × 28 pixels each [21];
- CIFAR10: ten classes of color images with 32 × 32 pixels each [20];
- PROTEIN: 357-dimensional sparse binary bio-samples in three classes [4];
- RCV1: 47, 236-dimensional sparse newswires from Reuters in two classes [22];
- SENSEIT: 100-dimensional samples from a vehicle net in three classes [8].

We implemented the calibration approaches in the MATLAB platform. For the cyclic calibration approach, the number of partitions was set to $r = 10$ unless otherwise specified. All results were recorded on a server with 28 CPU cores and 192GB memory enabled for computation.

4.2 Noise Reduction on Distance Metrics

One specific application scenario of the proposed approaches is noise reduction in given distance metrics. In each run of the experiment, we randomly chose 1, 000 samples from the MNIST dataset and computed their pairwise Euclidean distance matrix (D^*) as the ground truth metric. Next we added certain amounts of white noise to D^* and obtain a noisy metric D^0 with each $d^0_{ij} = \max\left\{0, d^*_{ij} + \zeta\mu v\right\}$, where μ is the mean of all elements in D^*, ζ was set to 0.1/0.5 respectively and $v \sim N(0, 1)$ is a standard Gaussian random variable.

We applied the direct calibration approach (denoted by DIRECT) and the cyclic calibration approach (denoted by CYCLIC) on D^0 and obtained two calibrated matrices ($D^0_{\mathcal{R}}$). The relative squared deviation (RSD) from D^*, calculated as $\frac{\left\|D^0_{\mathcal{R}}-D^*\right\|_F^2}{\left\|D^0-D^*\right\|_F^2}$, was recorded to measure the performance of each calibration method.

We repeated the experiment for ten runs and reported the mean of the results in Table 1. Compared with the noisy matrix D^0, the direct calibration reduced more than 86% of squared deviation when $\zeta = 0.1$ and more than 96% when $\zeta = 0.5$, and the cyclic calibration reported comparable improvements.

Table 2. Relative Squared Deviations on Calibration of Approximate Metrics from Incomplete Samples. Each item has two values, corresponding to $p = 0.1$ and 0.5 respectively: the smaller RSD value, the better performance. Direct calibration reported the best performances on almost all experiments.

	Imputation			Calibration		
Dataset	MEAN	kNN	SVT	TRIFIX	DIRECT	CYCLIC
MNIST	2.80/8.20	7.96/18.5	2.51/8.57	1.00/.998	**.998/.767**	1.12/.813
CIFAR10	765./243.	119./292.	25.1/61.5	1.00/1.00	**.991/.979**	1.37/.997
PROTEIN	53.0/17.7	1.77/3.40	1.79/3.12	.994/.924	**.975**/.506	1.05/**.464**
RCV1	1.48/3.00	1.52/3.07	1.44/2.89	.999/.931	**.806/.429**	.975/.430
SENSEIT	82.8/27.9	.908/1.13	.890/.861	.922/.502	**.874/.489**	.917/.543

We also recorded the performance of triangle fixing (TRIFIX) algorithm[2] (cf. Section 2.2), which calibrates the noisy metric to restore the triangle inequalities. The triangle fixing algorithm reduced around 2% and 64% squared deviations, respectively. As a comparison, our proposed approaches reported significantly superior calibration results. In addition to the MNIST dataset, we carried out the same experiment on the other datasets and found very similar results.

4.3 Distance Metrics from Incomplete Data

The second experiment was on estimating the distance metric from incomplete observations. In each of the ten runs, we randomly chose a subset of $1,000$ samples from the MNIST dataset and computed the pairwise distance matrix D^* as the ground truth.

Then, we randomly marked different portions ($p = 0.1/0.5$ respectively) of features as missing for each sample. For any two incomplete samples x_i and x_j in the dataset, denote $x_i(x_j)$ a new vector formed by keeping those features of x_i that are observed in both x_i and x_j. Based on the common features, an approximate distance for the two incomplete samples was given by:

$$d^0_{ij} = \|x_i(x_j) - x_j(x_i)\| \sqrt{\frac{q}{q_{ij}}}$$

where $q = 784$ is the dimension of the MNIST samples and q_{ij} is the number of features observed in both samples.

Let a distance matrix $D^0 = \left\{d^0_{ij}\right\}^n_{i,j=1}$. The matrix is often not embeddable, which leaves potential room for further calibration. Accordingly, we calibrated D^0 to a new estimate $D^0_{\mathcal{R}}$ by our proposed approaches, and computed their RSD values from the ground truth D^* as described in Section 4.2.

The two proposed approaches were compared with the triangle fixing algorithm (cf. Section 2.2) on the quality of the calibration. In addition, the results from

[2] Implementation downloaded from http://optml.mit.edu/software.html.

several imputation methods, which were popularly used in practice, were also included as a baseline. These imputation methods include:

- MEAN: Replacing missings by the observed mean of the feature;
- kNN: Replacing missings by weighted mean of $k = 5$ nearest samples [33];
- SVT: Low-rank matrix completion with singular value thresholding [3] [3].

We applied these imputation methods to replace the missing features with substituted values, calculated the distance matrix based on the imputed data, and recorded the corresponding RSD values. In the experiment, we also tested two implementations [26,13] of the classical expectation-maximization algorithm and the recent optimal transport algorithm [31] to impute the data. Unfortunately, different from their known excellent performances on low-dimensional data, both algorithms failed to execute on most of these high-dimensional data samples with a large portion of missing values. So their results were not available.

The results are given in Table 2. Compared with the un-calibrated D^0, we can see that the calibration approaches brought significant drops in squared deviations from the true D^*. Direct calibration reported the best results in RSD values on most of the datasets and the settings. When $p = 0.5$, it reduced around 23% to 57% squared deviations on most datasets. The only exception is on the CIFAR10 dataset, where the reduction of squared deviations is not that significant. However, the improvement from calibration approaches over the imputation methods is still significant.

At the same time, we can see that the imputation approaches had no guarantee of the quality of RSD values. The imputed data's distance matrix may be far from the ground truth. For example, naïvely filling the mean to the missing values on the CIFAR10 dataset produced a distance matrix that was more than seven hundred times away from the ground truth than that of D^0. Comparatively, the calibration approaches consistently reduced the squared deviation from the ground truth by calibrating the input matrix as expected.

4.4 Classification on Incomplete Samples

Having justified the capability of removing metric noises by the proposed approaches, we would like to investigate whether the calibrated results benefit real applications. Specifically, we applied the calibrated metrics in nearest neighbor classification tasks. Given a training set of samples with class labels, we tried to predict the labels of the samples in the testing set. For each testing sample, its label was predicted by the label of the nearest neighbor in the training set. Then the predicted label was compared against the accurate label to measure the classification performance.

We carried out one-versus-all cross-validation on the classification task. Each sample was used, in turn, as the testing sample, while all other samples formed the training set. We averaged all testing samples' classification errors and recorded the mean of average classification errors (MCE) over ten runs. Similar to the RSD

[3] Implementation downloaded from https://candes.su.domains/software/.

Table 3. Ten-Fold Mean Classification Errors on Incomplete Samples by Nearest Neighbor Classifier. Each item has two values, corresponding to $p = 0.1$ and 0.5 respectively: the smaller MCE values, the better performance. Direct calibration reported the best performances on almost all experiments.

		Imputation			Calibration		
	D^0	MEAN	kNN	SVT	TRIFIX	DIRECT	CYCLIC
MN	.127/.203	.145/.503	.132/.272	.150/.450	.127/.200	**.126**/.192	.127/**.191**
CI	.767/.765	.786/.878	.781/.842	.780/.838	.767/.765	**.751/.757**	.771/.758
PR	.581/.615	.604/.634	.638/.698	.595/.620	.581/.624	**.573/.610**	.574/.617
RC	.339/.442	.324/.432	.324/.435	.334/.437	.338/.463	**.321/.419**	.327/.427
SE	.299/.377	.394/.503	.299/.389	.292/.402	.292/.362	**.288/.357**	.301/.366

Table 4. Ten-Fold Mean Classification Errors on Incomplete Samples by Hard-Margin SVM with Gaussian Kernel and Default Parameters. One-versus-all strategy was applied for classifying more than two classes. Each item has two values, corresponding to $p = 0.1$ and 0.5 respectively: the smaller MCE values, the better performance. Direct calibration reported the best performances on almost all experiments.

		Imputation			Calibration		
	D^0	MEAN	kNN	SVT	TRIFIX	DIRECT	CYCLIC
MN	.100/.141	.105/.158	.098/.894	.105/.159	.100/.142	**.096/.136**	.097/.138
CI	.661/.664	.671/.694	.785/.903	.665/.691	.661/.664	**.656/.658**	.658/.670
PR	.516/.606	.399/.485	.401/.686	.400/.486	.516/.610	.452/.550	**.397/.484**
RC	.247/.370	.127/.287	.497/.523	.138/.258	.136/.384	**.124/.237**	.135/.248
SE	.373/.478	.239/.293	.249/.741	.240/.291	.369/.473	.264/.388	**.231/.289**

results in Table 2, the calibration approaches reported improved MCE results in Table 3. With different missing ratios $p = 0.1$ and $p = 0.5$, the calibration approaches consistently reduce the classification errors over the approximate metric D^0 on all datasets, among which the direct calibration approach performed the best. Comparatively, the metrics from imputation-based approaches sometimes performed even worse than D^0.

We further experimented with the support vector machines (SVM) algorithm [30]. SVM seeks a linear boundary with the maximum margin to separate two classes of samples in the feature space. To apply SVM, a positive semi-definite kernel matrix needs to be provided as the input to the algorithm. In the evaluation, we used the popular Gaussian kernel to construct the kernel matrix, $\exp\left(-\alpha D^2\right)$, where D^2 is the element-wise square of the metric obtained from each algorithm and α is the default kernel parameter set by the LibSVM package [4]. In case the kernel matrix constructed is not positive semi-definite (namely, D^0 and TRIFIX), a small positive number will be added to the diagonal elements to shift the matrix to be positive semi-definite. The one-versus-all strategy was applied for classifying more than two classes.

The MCE results are shown in Table 4. The proposed calibration methods reported similarly improved accuracies over the un-calibrated metric and the imputation approaches. The most significant improvement over the performance of D^0 was on the RCV1 dataset, from 0.247 to 0.124 and from 0.370 to 0.237 respectively. Consistent improvements were observed on the other datasets. Similarly, the calibration approaches reported superior results over the imputation approaches on most experiments. In the evaluation, we found that when the missing ratio is high, the performances of the imputation approaches become relatively unstable. For example, when $p = 0.5$, the misclassification error with kNN imputation significantly increased to 0.894 on the MNIST dataset and 0.903 on the CIFAR10 dataset, like a random guess. Comparatively, the calibration approaches' performances are much more reliable.

When comparing the proposed calibration approaches with the triangle fixing algorithm, we can find a similar trend of improvement in classification errors, although not as significant as the improvement over the imputation approaches. The improved classification accuracies are consistent with the results reported in Sections 4.2 and 4.3, which again justifies the benefits from the better-calibrated metrics to the unknown ground truth metric.

4.5 Scalability

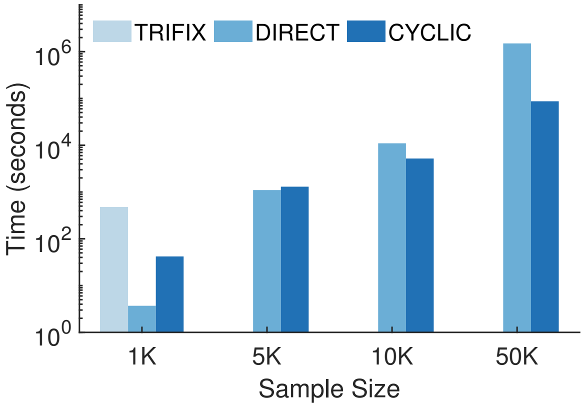

Fig. 1. Running Time on the MNIST Dataset with Different Sample Sizes. For $n = 1K/5K/10K$, $r = 10$; for $n = 50K$, $r = 50$. $|I_k| \approx \frac{2n}{r}$ $(1 \leq k \leq r)$. The triangle fixing algorithm failed to execute other than $n = 1K$. Cyclic calibration exhibited evidently improved scalability when the sample size is large.

The last experiment was to evaluate the scalability. Similarly to the setting described in Section 4.2, white noise was added to the factual distance matrix, and then the calibration approaches were carried out. Fig. 1 shows the running time in seconds of our two proposed approaches and the triangle fixing algorithm on the MNIST dataset with different numbers of training samples from $n = 1,000$ to $n = 50,000$.

With $n = 1,000$ training samples, the direct calibration approach took around three seconds, about a hundred times faster than the triangle fixing algorithm with the default parameter setting. When $n = 2,000$ (not shown in the figure) or larger, the triangle fixing algorithm failed to execute on our platform due to prohibitive memory requirement caused by the $O\left(n^3\right)$ triangle inequalities, so the results were not available here.

When comparing the direct and the cyclic calibration approaches, we can see that the cyclic approach did not report advantage with a small number of samples. However, when the number of samples got sufficiently large, e.g., $n = 10,000$, the cyclic approach began to exhibit its superiority. When $n = 50,000$, the cyclic approach was around twenty times faster than the direct approach to converge, being consistent with the complexity analysis in Section 3.5 and confirming a more scalable solution.

5 Conclusion

Estimating distance metrics between samples is a fundamental problem in data processing with various applications. To deal with the challenge, we suggested calibrating an approximate metric, which avoids the difficulty in imputation and returns an improved estimate with a solid guarantee. By connecting isometrical embedding and positive semi-definiteness of a distance matrix, the proposed approach provides a simple yet rigorous model for missing data processing, which forms the main contribution of our work. Computationally, Dykstra's alternating projection algorithm provides a natural solution to our proposed model and can be applied directly. Besides, our work also designed a cyclic projection algorithm that provided better scalability in the way of divide and conquer.

Compared with popular imputation methods, the proposed calibration approaches make fewer assumptions on the correlations among data features and the intrinsic data dimensions/ranks. As a result, the proposed approaches reported more reliable empirical results in our empirical evaluations of noise reduction and classification applications. Compared with existing models that can be applied for calibration purposes, such as the triangle fixing algorithm, the proposed approaches also reported significantly improved speed and accuracy. Although preliminary, all the results clearly justified the proposed approaches' benefits and demonstrated their high potential in practical tasks.

Despite the achieved results, more work along this line deserves to be investigated. The improved performance of our work relies on the assumption that the data samples can be isometrically embeddable in a Hilbert/Euclidean space. However, this assumption may not hold for general metrics. For example, the

Robinson-Foulds distance metric [32] defined on trees satisfies the triangle inequalities but is typically not embeddable. Can we extend the proposed approach to calibrate such metrics? It deserves our investigation. Another potential topic, although the cyclic calibration approach exhibited better scalability, it still seems demanding when handling big data, and the scalability issue deserves further consideration.

Acknowledgments

We thank the reviewers for the helpful comments. The work is supported by Guangdong Basic and Applied Basic Research Foundation (2021A1515011825), Shenzhen Science and Technology Program (CUHKSZWDZC0004), and Shenzhen Research Institute of Big Data.

References

1. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, New York, NY, USA (2004)
2. Brickell, J., Dhillon, I., Sra, S., Tropp, J.: The metric nearness problem. SIAM Journal on Matrix Analysis and Applications **30**(1), 375–396 (2008)
3. Cai, J.F., Candès, E., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization **20**(4), 1956–1982 (2010)
4. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**(3), 1–27 (2011)
5. Cline, A., Dhillon, I.: Computation of the singular value decomposition. In: Handbook of Linear Algebra, pp. 45–1. Chapman and Hall/CRC (2006)
6. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39**, 1–38 (1977)
7. Deutsch, F.: Best Approximation in Inner Product Spaces. Springer, New York, NY, USA (2001)
8. Duarte, M., Hu, Y.: Vehicle classification in distributed sensor networks. Journal of Parallel and Distributed Computing **64**(7), 826–838 (2004)
9. Duda, R., Hart, P.: Pattern Classification. John Wiley and Sons, Hoboken, NJ, USA (2000)
10. Dykstra, R.: An algorithm for restricted least squares regression. Journal of the American Statistical Association **78**(384), 837–842 (1983)
11. Enders, C.: Applied Missing Data Analysis. Guilford Press (2010)
12. Escalante, R., Raydan, M.: Alternating Projection Methods. SIAM, Philadelphia, PA, USA (2011)
13. Ghahramani, Z., Jordan, M.: Supervised learning from incomplete data via an EM approach. Advances in Neural Information Processing Systems **6**, 120–127 (1994)
14. Gilbert, G.: Positive definite matrices and Sylvester's criterion. The American Mathematical Monthly **98**(1), 44–46 (1991)
15. Golub, G., Van Loan, C.: Matrix Computations. Johns Hopkins University Press, Baltimore, MD, USA (1996)
16. Higham, N.: Computing the nearest correlation matrix - a problem from finance. IMA Journal of Numerical Analysis **22**, 329–343 (2002)

17. Horn, R., Johnson, C.: Matrix Analysis. Cambridge University Press (2012)
18. Jain, A., Murty, M., Flynn, P.: Data clustering: A review. ACM Computing Surveys **31**(3), 264–323 (1999)
19. Jannach, D., Resnick, P., Tuzhilin, A., Zanker, M.: Recommender systems—beyond matrix completion. Communications of the ACM **59**(11), 94–102 (2016)
20. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
21. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
22. Lewis, D., Yang, Y., Rose, T., Li, F.: RCV1: A new benchmark collection for text categorization research. Journal of Machine Learning Research **5**(Apr), 361–397 (2004)
23. Li, W.: Estimating Jaccard index with missing observations: a matrix calibration approach. Advances in Neural Information Processing Systems **28**, 2620–2628 (2015)
24. Li, W.: Scalable calibration of affinity matrices from incomplete observations. In: Asian Conference on Machine Learning. pp. 753–768 (2020)
25. Little, R., Rubin, D.: Statistical analysis with missing data, vol. 793. John Wiley & Sons (2019)
26. Murphy, K.: Machine Learning: a Probabilistic Perspective. MIT Press (2012)
27. Muzellec, B., Josse, J., Boyer, C., Cuturi, M.: Missing data imputation using optimal transport. In: International Conference on Machine Learning. pp. 7130–7140. PMLR (2020)
28. Qi, H., Sun, D.: An augmented Lagrangian dual approach for the H-weighted nearest correlation matrix problem. IMA Journal of Numerical Analysis **31**(2), 491–511 (2011)
29. Schoenberg, I.: Metric spaces and positive definite functions. Transactions of the American Mathematical Society **44**(3), 522–536 (1938)
30. Schölkopf, B., Smola, A., Bach, F., et al.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press (2002)
31. Sonthalia, R., Gilbert, A.C.: Project and forget: Solving large-scale metric constrained problems. arXiv preprint arXiv:2005.03853 (2020)
32. Stockham, C., Wang, L.S., Warnow, T.: Statistically based postprocessing of phylogenetic analysis by clustering. Bioinformatics **18**(suppl_1), S285–S293 (2002)
33. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, R., et al.: Missing value estimation methods for DNA microarrays. Bioinformatics **17**(6), 520–525 (2001)
34. Wells, J., Williams, L.: Embeddings and Extensions in Analysis, vol. 84. Springer Science & Business Media (1975)
35. Xing, E., Jordan, M., Russell, S., Ng, A.: Distance metric learning with application to clustering with side-information. Advances in Neural Information Processing Systems **15**, 521–528 (2002)