Article

Relational Fisher Analysis: Dimensionality Reduction in Relational Data with Global Convergence †

Li-Na Wang [1], **Guoqiang Zhong** [2,*] , **Yaxin Shi** [2] **and Mohamed Cheriet** [3]

[1] Qingdao Vocational and Technical College of Hotel Management, Qingdao 266100, China; alinagq@163.com
[2] College of Computer Science and Technology, Ocean University of China, Qingdao 266100, China; syxzhuanshu@163.com
[2] Synchromedia Laboratory for Multimedia Communication in Telepresence, École de Technologie Supérieure, Montréal, QC H3C 1K3, Canada; mohamed.cheriet@etsmtl.ca
[*] Correspondence: gqzhong@ouc.edu.cn
[†] This paper is an extended version of our paper published in Zhong, G.; Shi, Y.; Cheriet, M. Relational Fisher analysis: A general framework for dimensionality reduction. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016.

Abstract: Most of the dimensionality reduction algorithms assume that data are independent and identically distributed (i.i.d.). In real-world applications, however, sometimes there exist relationships between data. Some relational learning methods have been proposed, but those with discriminative relationship analysis are lacking yet, as important supervisory information is usually ignored. In this paper, we propose a novel and general framework, called relational Fisher analysis (RFA), which successfully integrates relational information into the dimensionality reduction model. For nonlinear data representation learning, we adopt the kernel trick to RFA and propose the kernelized RFA (KRFA). In addition, the convergence of the RFA optimization algorithm is proved theoretically. By leveraging suitable strategies to construct the relational matrix, we conduct extensive experiments to demonstrate the superiority of our RFA and KRFA methods over related approaches.

Keywords: relational learning; dimensionality reduction; graph embedding; trace ratio; document understanding and recognition; face recognition

1. Introduction

In some applications, such as pattern recognition and data mining, dimensionality reduction methods are often used since they can reduce space-time complexity, denoise, and make the model more robust. Principal component analysis (PCA) [1–3] and linear discriminant analysis (LDA) [2,4–7] are two typical linear algorithms. Following them, researchers have proposed many variants, such as kernel PCA [8], generalized discriminant analysis (GDA) [9], and linear discriminant analysis for robust dimensionality reduction (RLDA) [10].

For nonlinear dimensionality reduction problems, manifold learning provides an effective solution. By supposing that data are located on a low-dimensional manifold, data samples observed in high-dimensional space can be represented in a low-dimensional space. Some representative manifold learning algorithms are ISOMAP [11], locally linear embedding (LLE) [12] and Laplacian Eigenmaps (LE) [13].

From the algorithmic perspective, algorithms mentioned above can be categorized as global methods or local methods. Global methods learn the low-dimensional representations by using global information of data. PCA and LDA are all global methods. The global methods are often effective and efficient, such that they are widely used in many real world applications. However, when dealing with non-linear data, using the global method cannot capture the genuine distribution of data very well. Local methods using the manifold learning idea, such as LLE and LE, pay special attention to the intrinsic

structure of data. Nevertheless, most of these manifold learning methods disregard label information when recovering the low-dimensional manifold structure, in that they are inherently unsupervised.

Although the above-mentioned methods are defined from different perspectives of dimensionality reduction, graph embedding provide a unified framework for understanding and comparing them [14]. Furthermore, by integrating label information in the computation of intrinsic and penalty graphs within the graph embedding framework, a supervised dimensionality reduction method called marginal Fisher analysis (MFA) is proposed [14].

In traditional dimensionality reduction algorithms as described above, a data distribution assumption is generally applied that data are independent and identically distributed (i.i.d.). However, in real-world applications, there are often certain relativity or links between certain data, for instance, geometrical or semantic similarity, links among web pages, citation relations between scientific papers. Relationships usually indicate that these related samples are likely to have similarities or belonging to the same class. Nevertheless, although some dimensionality reduction methods consider to preserve the locality of data [15–18], the useful relationships are often simply ignored during the learning process of most existing dimensionality reduction methods.

Recently, relational learning has often been used in practical applications, for instance, web mining [19] and social network analysis [20]. In addition, relational information is also considered in social network discovery, document classification, sequential data analysis and semi-supervised graph embedding [21–23].

In the domain of dimensionality reduction, some algorithms have already been proposed in which the relational information is integrated into the representation learning process. In [24], Duin et al. propose the relational discriminant analysis (RDA). In RDA, relationships among data are measured by the Euclidean distance between objects and prototypes or support objects of each class. However, RDA uses mean squared error as the objective function, and cannot perform well on multi-class learning problems. In [25], Li et al. propose the probabilistic relational PCA (PRPCA) to build a probabilistic model associated with PCA and relational learning. RDA and PRPCA effectively integrate relational learning into the dimensionality reduction algorithms. Nevertheless, how to better inject relationships into traditional dimensionality reduction models is still worth exploring.

Recently, a few deep relational learning algorithms have been proposed. Specifically, Gao et al. design a deep learning model based on relational network for hyperspectral image few-shot classification [26], Chen et al. apply local relation learning for face forgery detection [27], and Cho et al. develop a weakly supervised anomaly detection method via context-motion relational learning [28]. In addition, some relational learning methods are used in the one-shot [29] or zero-shot [30] learning scenarios. However, many real-world applications have only very few data. Hence, shallow relational learning algorithms are still needed to be proposed and utilized.

In this paper, we propose a novel and general framework for dimensionality reduction, called relational Fisher analysis (RFA) [31]. Besides the intrinsic and penalty graph in the graph embedding framework, we further construct a relational graph which captures the relational information encoded within data. Through this graph, the proposed RFA takes into account the impact of the relational information in the presentation learning process. An effective iterative trace ratio algorithm is proposed to optimize RFA. Futhermore, we use the kernel trick to extend RFA to its kernelized version—KRFA. Additiionally, we theoretically prove that the optimization algorithm of RFA converges. To evaluate the effectiveness of RFA, we conduct extensive experiments in many real-world applications. The results demonstrate that the proposed RFA outperforms most of the classic dimensionality reduction algorithms on the datasets we use. The effectiveness of KRFA is also tested.

This paper is based on one of our previous conference papers [31], with significant improvements. For concreteness, we propse the KRFA algorithm and add more exclusive experiments with comparison to the related approaches. The rest of this paper is organized as follows: In Section 2, we introduce several related ideas, including graph embedding,

trace ratio problem and relational learning, which are highly relevant to our work. In Section 3, we focus on our proposed method RFA, including the notation, the formulation and the iterative optimization method of RFA. Section 4 includes the proof of the convergence of RFA and we present how to extend RFA to a kernel version—KRFA. In Section 5, we compare our methods RFA and KRFA to other commonly used dimensionality reduction methods with extensive experiments, which demonstrate the effectiveness of our proposed methods. Finally, we summarize this paper in Section 6.

2. Related Work

In this section, we first briefly introduce some traditional dimensionality reduction methods, and then offer a detailed description about relevant ideas including graph embedding, trace ratio problem and relational learning, respectively. Finally, we specify how those ideas are used in this work.

2.1. Traditional Dimensionality Reduction Methods

Some basic ideas of traditional dimensionality reduction methods are presented in this subsection, such as PCA, LDA as well as several locality based manifold learning methods including LLE and LE. Advantages and drawbacks of these methods are also presented in this part.

2.1.1. PCA

The main idea of PCA [1] is to seek projection directions with maximal variances of the low-dimensional embeddings. It effectively extract and retain the principle components of the original data. However, as PCA is an unsupervised dimensionality reduction method, low-dimensional embeddings obtained from this method cannot perfectly maintain the discrimination between data of different classes.

2.1.2. LDA

LDA [7], well known as a supervised dimensionality reduction method, aims to seek projection directions to minimize the intraclass scattering and maximize the interclass scattering for the low-dimensional embeddings. However, for LDA, if the dimensionality of data is far greater than the data size, the intraclass scattering matrix may suffer from the singularity problem and thus it influences the solution of this dimensionality reduction algorithm. Furthermore, since the rank of the interclass scattering matrix is at most $C - 1$, the number of available projection directions of LDA is at most $C - 1$, where C is the number of classes.

2.1.3. Manifold Learning Methods

PCA and LDA are all global methods which use global information to project the original data into a subspace and obtain the low-dimensional data representations. However, for highly nonlinear data structure, these linear methods cannot learn the nonlinear relationships between data and thus the results are not ideal. By assuming that the high-dimensional data have a low-dimensional manifold structure, manifold learning algorithms can nonlinearly map the high-dimensional data onto their low-dimensional manifold. Among manifold learning methods, local geometric information-based methods, such as LLE, LE and local preserving projection (LPP) [32] are widely used. The ideas behind them are as follows.

LLE [12] preserves the linear reconstruction characteristics in a local neighborhood of each datum. Hence, the low-dimensional embeddings obtained by LLE presents the local geometrical structure of the data manifold. LE [13] preserves the similarities of the neighboring data points based on an adjacency matrix and a graph Laplacian matrix. However, for LLE and LE, as the nonlinear mapping function between the high-dimensional and low-dimensional spaces is not learned, we cannot easily obtain the low-dimensional

representations of new data. To the end, LPP [32] performs a linear approximation of LE, and successfully overcomes its drawback as mentioned above.

2.2. Graph Embedding

In [14], Yan et al. show that some commonly used dimensionality reduction algorithms could be transformed into a unified framework despite their different motivations, and the unified framework is called graph embedding. This framework derives a low-dimensional feature space, which preserves the adjacency relationship between sample pairs. The general objective function of this framework is presented in Equation (1), where $\mathbf{W}$ denotes the similarity matrix of the undirected weighted graph $\mathbf{G} = \{\mathbf{X}, \mathbf{W}\}$ and $\mathbf{B}$ is the constraint matrix defined to avoid a trivial solution of the objective function,

$$\mathbf{V}^* = \underset{\mathbf{V}^T \mathbf{S}_B \mathbf{V} = \mathbf{I}}{\arg\min} \mathbf{V}^T \mathbf{S}_W \mathbf{V} = \underset{\mathbf{V}}{\arg\min} \frac{\mathbf{V}^T \mathbf{S}_W \mathbf{V}}{\mathbf{V}^T \mathbf{S}_B \mathbf{V}}, \tag{1}$$

where $\mathbf{S}_W$ and $\mathbf{S}_B$ are matrices constructed with respect to $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{B}$, respectively.

We note that this unified framework graph embedding also provides a new idea for researchers to propose new dimensionality reduction algorithms. In particular, Yan et al. propose a novel dimensionality reduction method by defining an intrinsic graph which characterizes the intraclass compactness and a penalty graph which characterizes the interclass separability in the graph embedding framework, and call it marginal Fisher analysis (MFA).

2.3. Trace Ratio Problems

As presented in the above subsection, within the context of graph embedding, the dimensionality reduction methods can be viewed as trying to obtain the transformation matrix $\mathbf{W}$ that makes $Tr(\mathbf{W}^T \mathbf{S}_p \mathbf{W})$ maximum and $Tr(\mathbf{W}^T \mathbf{S}_l \mathbf{W})$ minimum. This is often formulated as a trace ratio optimization problem, that is $\max_{\mathbf{W}} Tr(\mathbf{W}^T \mathbf{S}_p \mathbf{W}) / Tr(\mathbf{W}^T \mathbf{S}_l \mathbf{W})$ [33]. Generally, there are two kinds of solutions for this problem: (1) Simplifying the problem into a ratio trace problem: $\max_W Tr[(\mathbf{W}^T S_l \mathbf{W})^{-1}(\mathbf{W}^T S_p \mathbf{W})]$, then using generalized eigenvalue decomposition (GED) to obtain the transformation matrix $\mathbf{W}$; (2) Directly optimizing the objective function through an iterative procedure, with each step presented as a trace difference problem: $Tr[(\mathbf{W}^T (\mathbf{S}_p - \lambda^n \mathbf{S}_l) \mathbf{W})]$. However, for the first solution, the optimization of ratio trace formulation may deviate from the original objective, which results in a closed-form but inexact solution and may subsequently lead to uncertainty in subsequent classification or clustering problems. For the second solution, Wang et al. [33] propose an efficient iterative procedure by solving the trace difference problem in each iterative step, named iterative procedure (ITR). It is proven that ITR could converge to the optimal solution and solve the trace ratio problem. With the orthogonal assumption on the projection matrix, objective function of the ITR optimization can be defined as

$$\mathbf{W} = \underset{\mathbf{W}^T \mathbf{W} = \mathbf{I}}{\arg\max} \frac{\mathrm{tr}(\mathbf{W}^T \mathbf{S}_p \mathbf{W})}{\mathrm{tr}(\mathbf{W}^T \mathbf{S}_t \mathbf{W})}. \tag{2}$$

In [34], Nie et al. address the graph-based feature selection framework using the iterative process of the trace ratio problem. In [35], Zhong et al. analyze the iterative procedures for the trace ratio problem and prove necessary and sufficient conditions of the existence of the optimal solution of trace ratio problems, which are that there is a sequence $\{\lambda_1^*, \lambda_2^*, \dots, \lambda_n^*\}$ that converges to λ^* as $n \to +\infty$, where λ^* is the optimal value of Equation (2). Based on these previous works, we also formulate RFA as a trace ratio problem and theoretically prove the convergence of its optimization algorithm.

2.4. Relational Learning

In many real-world applications, data generally share some kinds of relations, such as geometrical or semantic similarity, links or citations. This relation information encoded

inside data provides valuable evidence for some issues, such as classification and retrieval. To the end, relational learning is generally integrated into the representation learning models.

In [36], Duin et al. prove that it is possible to use only proximity measure (distances or similarities) to represent the samples rather than mapping the feature vectors to the low-dimensional space. In addition, they propose a proximity description-based dimensionality reduction method called relational discriminant analysis (RDA) in [24]. Instead of data, RDA uses similarities to a subset of objects in the training data as features. In this case, dimensionality reduction can be conducted either by selection methods (such as random selection [37], systematic selection [38]), or by feature extraction methods (such as multi-dimensional scaling [39], Sammon mapping [40] and Niemann mapping [41]).

In [25], Li et al. model the covariance of data with the relationships between instances and propose a Gaussian latent variable model which successfully integrates relational information into the dimensionality reduction process, called probabilistic relational PCA (PRPCA). In PRPCA, relational information is defined by the relevance between data samples. We take the scientific paper citation as an example. If there is a quoting between the papers, it means that these papers most likely have similar topics. To take the inter-influence between cited papers into account, Li et al. further construct a matrix $\mathbf{\Phi} = \mathbf{\Delta}^{-1}$, which satisfy the condition that similar instances often have a lower probability density at the latent space. To the end, PRPCA, based on the relational covariance $\mathbf{\Phi}$, successfully applies the relational information to the dimensionality reduction algorithms.

Relational learning is also commonly used for data mining, information retrieval and other machine learning-related applications. Paccanaro et al. [42] propose a method, called linear relational embedding, for the distributed representations of data, where data consist of the relationship of concepts. Wang et al. [43] utilize the characteristic that existing relations between items are often useful in recommendation systems and propose a model called relational collaborative topic regression (RCTR), which expand the traditional CTR model by integrating feedback information, item content information and relational information. Xuan et al. [44] propose a nonparametric relational topic model using stochastic processes instead of fixed-dimensional probability distributions.

Based on the classical works mentioned above, we propose a general and effective dimensionality reduction framework named relational Fisher analysis (RFA). This framework uses graph embedding [14] as theoretical foundation and integrates the relational information [24,25] encoded inside data into the dimensionality reduction process. Besides the intrinsic graph and the penalty graph as defined using graph embedding, we further construct a relational graph based on the existing relationships between data, which enables the desired low-dimensional space to preserve the intrinsic information, reduce the penalty information and further learn and preserve the relational information among the data samples. In addition, through the derivation and equivalent transformation operations, the objective function of our proposed method can be transformed into the trace ratio form for optimization. Based on a systematic analysis of two optimization method for trace ratio problems [33], we propose a novel iterative algorithm which uses the value of the trace ratio as criterion for the algorithmic convergence. In addition, by further introducing the ITR-Score defined in [34] into the iterative process, optimal projection directions are learned, which improves the effectiveness of the proposed RFA model.

3. Methodology

In this section, we first present some notations used in our work. The iterative steps, the optimization method and the proof of global convergence of RFA are then introduced in detail.

3.1. Notation

Matrices are represented in uppercase bold letters, for instance, $\mathbf{A}$, while vectors are represented in boldface lowercase letters, for instance, $\mathbf{a}$, and $\mathbf{a}_i$ is the ith element of $\mathbf{a}$. $\mathbf{A}_{i*}$ and $\mathbf{A}_{*j}$ denote the ith row and jth column of a matrix $\mathbf{A}$; therefore, the element of the

*i*th row and *j*th column of the matrix is represented by $\mathbf{A}_{ij}$. The trace of $\mathbf{A}$ is defined by $\mathrm{tr}(\mathbf{A})$ and the transpose of $\mathbf{A}$ is defined by $\mathbf{A}^T$. In addition, $|\mathbf{A}_{ij}|$ is the absolute value of $\mathbf{A}_{ij}$, $\|\mathbf{A}\|_F$ is the Frobenius norm of $\mathbf{A}$. If $\mathbf{A}$ is positive definite, we have $\mathbf{A} \succ 0$, while it is positive semi-definite (psd), we have $\mathbf{A} \succeq 0$.

In a learning task that contains multiple classes of data, we usually have dataset $\{\{\mathbf{X}_{*i}, \mathbf{y}_i\} \in \Re^D \times \Re^1, i = 1, 2, \ldots, N\}$, where each $\mathbf{X}_{*i}$ represents a sample and $\mathbf{y}_i \in \{1, 2, \ldots, C\}$ is the class of that sample, $C \geqslant 2$ is the total number of classes and N is the total number of samples. For a linear dimensionality reduction task, we hope to find a projection matrix $\mathbf{W}$ and obtain the *d*-dimensional representation $\mathbf{Z}_{*i}$ of $\mathbf{X}_{*i}$ by $\mathbf{Z}_{*i} = \mathbf{W} * \mathbf{X}_{*i}$, where $\mathbf{Z}_{*i} \in \Re^d, i = 1, 2, \ldots, N, d < D$ is the dimensionality of the output.

3.2. Formulation of RFA

As discussed in Section 2, graph embedding has already been proven to be a general framework for dimensionality reduction. However, there are two shortages of graph embedding. First, graph embedding does not obtain and preserve the relational information between data. Second, the graph embedding framework needs to be solved by generalized eigenvalue decomposition, which is only an approximate approach. Inspired by graph embedding, we propose a new dimensionality reduction framework called RFA, which integrates relationships among data into the dimensionality reduction model and can alleviate the two problems mentioned above. Formulation of the proposed RFA is described as follows.

We use $\mathbf{R} \in \Re^{N \times N}$ to denote the relational matrix. The dimensionality reduction framework RFA is modeled as

$$\mathcal{L} = \min_{\mathbf{W}} \frac{\mathrm{tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}_I \mathbf{X}^T \mathbf{W})}{\mathrm{tr}(\mathbf{W}^T \mathbf{X} \mathbf{L}_P \mathbf{X}^T \mathbf{W})} + \lambda \mathrm{tr}(\mathbf{W}^T \mathbf{X} \mathbf{R} \mathbf{X}^T \mathbf{W}), \tag{3}$$

where $\mathbf{L}_I$ and $\mathbf{L}_P$ define the intrinsic and penalty graphs, respectively, and $\lambda \geqslant 0$ is a hyperparameter. Specifically, we only consider undirected graph and assume that $\mathbf{L}_I$, $\mathbf{L}_P$ and $\mathbf{R}$ are symmetric and psd. Based on this formulation and these assumptions, the generality of RFA can be explained from the following two points:

(1) If $\lambda = 0$, our algorithm can be simplified to a basic graph embedding model, so that some commonly used dimensionality reduction algorithms can be regarded as special cases of RFA;

(2) Otherwise, if $\mathcal{L}$ only contains relational information, RFA can be considered to use relational learning to reduce the dimensionality of data. For instance, the MDS algorithm is a special RFA algorithm under this condition.

3.3. Optimization of RFA

We reformulate Problem (3) as

$$\mathcal{L} = \min_{\mathbf{W}} \frac{\mathrm{tr}(\mathbf{W}^T \mathbf{S}_I \mathbf{W})}{\mathrm{tr}(\mathbf{W}^T \mathbf{S}_P \mathbf{W})} + \lambda \mathrm{tr}(\mathbf{W}^T \mathbf{S}_R \mathbf{W}), \tag{4}$$

where $\mathbf{S}_I = \mathbf{X} \mathbf{L}_I \mathbf{X}^T$, $\mathbf{S}_P = \mathbf{X} \mathbf{L}_P \mathbf{X}^T$ and $\mathbf{S}_R = \mathbf{X} \mathbf{R} \mathbf{X}^T$.

As $\mathbf{R}$ is psd, $\mathbf{S}_R$ is as well. We suppose $\mathbf{S}_R = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$. We have

$$\mathcal{L} = \min_{\mathbf{V}} \frac{\mathrm{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_I \mathbf{V})}{\mathrm{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V})} + \lambda \mathrm{tr}(\mathbf{V}^T \mathbf{V}), \tag{5}$$

where $\mathbf{W} = \mathbf{U} \mathbf{\Lambda}^{-1/2} \mathbf{V}$, $\widetilde{\mathbf{S}}_I = \mathbf{\Lambda}^{-1/2} \mathbf{U}^T \mathbf{S}_I \mathbf{U} \mathbf{\Lambda}^{-1/2}$ and $\widetilde{\mathbf{S}}_P = \mathbf{\Lambda}^{-1/2} \mathbf{U}^T \mathbf{S}_P \mathbf{U} \mathbf{\Lambda}^{-1/2}$.

Furthermore,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = \frac{2\mathrm{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V}) \widetilde{\mathbf{S}}_I \mathbf{V} - 2\mathrm{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_I \mathbf{V}) \widetilde{\mathbf{S}}_P \mathbf{V}}{\mathrm{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V})^2} + 2\lambda \mathbf{V}. \tag{6}$$

We let $\frac{\partial \mathcal{L}}{\partial \mathbf{V}} = \mathbf{0}$. We have

$$\frac{\text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V}) \widetilde{\mathbf{S}}_I \mathbf{V} - \text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_I \mathbf{V}) \widetilde{\mathbf{S}}_P \mathbf{V}}{\text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V})^2} = -\lambda \mathbf{V}. \tag{7}$$

Equation (7) can be rewritten as

$$(\widetilde{\mathbf{S}}_I - \frac{\text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_I \mathbf{V})}{\text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V})} \widetilde{\mathbf{S}}_P) \mathbf{V} = -\lambda \text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V}) \mathbf{V}. \tag{8}$$

We let $\eta = \frac{\text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_I \mathbf{V})}{\text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V})}$ and $\widetilde{\lambda} = -\lambda \text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V})$. We obtain

$$(\widetilde{\mathbf{S}}_I - \eta \widetilde{\mathbf{S}}_P) \mathbf{V} = \widetilde{\lambda} \mathbf{V}. \tag{9}$$

It can be seen from Equation (9) that the columns of matrix $\mathbf{V}$ are the eigenvectors of $\widetilde{\mathbf{S}}_I - \eta \widetilde{\mathbf{S}}_P$, where η is a parameter related to $\mathbf{V}$.

Without loss of generality, we assume $\mathbf{V}^T \mathbf{V} = \mathbf{I}_d$, where $\mathbf{I}_d$ is an identity matrix. Hence, we have the following constrained trace ratio problem [33,45]:

$$\widetilde{\mathcal{L}} = \min_{\mathbf{V}^T \mathbf{V} = \mathbf{I}_d} \frac{\text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_I \mathbf{V})}{\text{tr}(\mathbf{V}^T \widetilde{\mathbf{S}}_P \mathbf{V})}. \tag{10}$$

Problem (10) can be solved with the iterative method similar to that in [33,45]. The specific steps are as follows:

(1) **Removing the null space of $\mathbf{S}_t = \widetilde{\mathbf{S}}_I + \widetilde{\mathbf{S}}_P$ [46].** We assume that $\mathbf{S}_t = \widetilde{\mathbf{U}} \widetilde{\mathbf{\Lambda}} \widetilde{\mathbf{U}}^T$, where $\widetilde{\mathbf{\Lambda}}$ is a diagonal matrix and $\widetilde{\mathbf{U}}$ contains the eigenvectors of $\mathbf{S}_t$ corresponding to nonzero eigenvalues. Therefore, Formula (10) can be transformed to

$$\widetilde{\mathcal{L}} = \min_{\widetilde{\mathbf{W}}^T \widetilde{\mathbf{W}} = \mathbf{I}_d} \frac{\text{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}})}{\text{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_P \widetilde{\mathbf{W}})}, \tag{11}$$

where $\mathbf{V} = \widetilde{\mathbf{U}} \widetilde{\mathbf{W}}, \widetilde{\mathbf{W}} \in \Re^{r \times d}, \widehat{\mathbf{S}}_I = \widetilde{\mathbf{U}}^T \widetilde{\mathbf{S}}_I \widetilde{\mathbf{U}}$ and $\widehat{\mathbf{S}}_P = \widetilde{\mathbf{U}}^T \widetilde{\mathbf{S}}_P \widetilde{\mathbf{U}}$. We can further rewrite the problem (11) as

$$\widetilde{\mathcal{L}} = \min_{\widetilde{\mathbf{W}}^T \widetilde{\mathbf{W}} = \mathbf{I}_d} \frac{\text{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}})}{\text{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_T \widetilde{\mathbf{W}})}, \tag{12}$$

where $\widehat{\mathbf{S}}_T = \widetilde{\mathbf{U}}^T (\widetilde{\mathbf{S}}_I + \widetilde{\mathbf{S}}_P) \widetilde{\mathbf{U}} = \widehat{\mathbf{S}}_I + \widehat{\mathbf{S}}_P$. Since $\widehat{\mathbf{S}}_T$ is positive definite, for any orthonormal matrix $\widetilde{\mathbf{W}}$, Problem (12) satisfies that the denominator is positive.

(2) **Efficient iterative optimization.** The original trace ratio problem (12) can be rewritten as a trace difference problem:

$$\widetilde{\mathbf{W}}^* = \underset{\widetilde{\mathbf{W}}^T \widetilde{\mathbf{W}} = \mathbf{I}_d}{\text{argmin}} \, \text{tr}(\widetilde{\mathbf{W}}^T (\widehat{\mathbf{S}}_I - \widetilde{\eta} \widehat{\mathbf{S}}_T) \widetilde{\mathbf{W}}), \tag{13}$$

where $\widetilde{\eta}$ is a parameter which can be calculated in the iterative process. In the iterative process, we first randomly initialize the target matrix $\widetilde{\mathbf{W}}$ to be an arbitrary orthogonal matrix as $\widetilde{\mathbf{W}}_0 \in \Re^{r \times d}$, and then calculate $\widetilde{\eta}_0 = \frac{\text{tr}(\widetilde{\mathbf{W}}_0^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}}_0)}{\text{tr}(\widetilde{\mathbf{W}}_0^T \widehat{\mathbf{S}}_T \widetilde{\mathbf{W}}_0)}$. By using the calculated $\widetilde{\eta}_0$, we can obtain $\widetilde{\mathbf{W}}_1$ by solving Problem (13). In the end, through several iterations, we obtain $\widetilde{\mathbf{W}}_T$, where T is the number of iterations and it's satisfied $|\widetilde{\eta}_T - \widetilde{\eta}_{T-1}| < \epsilon$ ($\epsilon = 10^{-5}$ is used in our experiments). Then, $\widetilde{\mathbf{W}}_T$ is the optimal solution of Problem (12). In next section, we prove that RFA owns the global convergence. In order to improve the effectiveness of our

method, we select some superior projection directions for each $\widetilde{\mathbf{W}}_t$, as performed in [47]. Our selection criterion is

$$\widetilde{\mathbf{W}}_t = \underset{\widetilde{\mathbf{W}} \in \Phi}{\operatorname{argmin}} \frac{\operatorname{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}})}{\operatorname{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_T \widetilde{\mathbf{W}})}, \tag{14}$$

where Φ is a set of $r \times d$ matrices with columns formed by eigenvectors of $\widehat{\mathbf{S}}_I - \widetilde{\eta}_{t-1}\widehat{\mathbf{S}}_T$. We use the eigenvectors corresponding to d smallest ITR-score [48] to initialize the selection.

Algorithm 1 specifically describes the iterative procedure of Problem (12).

Algorithm 1 Optimization of Problem (12)

1: **Initialization:** Initialize $\widetilde{\mathbf{W}}$ as an orthonormal matrix $\widetilde{\mathbf{W}}_0 \in \Re^{r \times d}$ and Let $\widetilde{\eta}_0 = 0$.
2: **Iterations:**
3: **for** $t = 0$ **to** *MaxIt* **do**
4:     (1) Compute $\widetilde{\eta}_t$ as

$$\widetilde{\eta}_t = \frac{\operatorname{tr}(\widetilde{\mathbf{W}}_{t-1}^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}}_{t-1})}{\operatorname{tr}(\widetilde{\mathbf{W}}_{t-1}^T \widehat{\mathbf{S}}_T \widetilde{\mathbf{W}}_{t-1})}.$$

5:     (2) Solve the eigenvalue decomposition problem:

$$(\widehat{\mathbf{S}}_I - \widetilde{\eta}\widehat{\mathbf{S}}_T)\widehat{\mathbf{W}}_{*i} = \widetilde{\lambda}_i\widehat{\mathbf{W}}_{*i}.$$

6:     (3) Compute

$$s_i = \frac{\operatorname{tr}(\widehat{\mathbf{W}}_{*i}^T \widehat{\mathbf{S}}_I \widehat{\mathbf{W}}_{*i})}{\operatorname{tr}(\widehat{\mathbf{W}}_{*i}^T \widehat{\mathbf{S}}_T \widehat{\mathbf{W}}_{*i})}, \ i = 1, \dots, r.$$

7:     (4) Use $\{\widehat{\mathbf{W}}_{*i}\}_{i=1}^r$ to initialize $\widetilde{\mathbf{W}}_t$ and solve the following problem:

$$\widetilde{\mathbf{W}}_t^* = \underset{\widetilde{\mathbf{W}} \in \Phi}{\operatorname{argmin}} \frac{\operatorname{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}})}{\operatorname{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_T \widetilde{\mathbf{W}})},$$

    where Φ is a set of matrices with columns formed by $\{\widehat{\mathbf{W}}_{*i}\}_{i=1}^r$.
8:     **if** $|\widetilde{\eta}_t - \widetilde{\eta}_{t-1}| < \epsilon$ ($\epsilon = 10^{-5}$ is used in our experiments) **then**
9:         Break.
10:     **end if**
11: **end for**
12: **Output:** $\widetilde{\mathbf{W}}_t$.

---

4. Global Convergence of RFA and Extensions

In this section, we first prove that RFA can converge to the global optimal solution, and then, we apply the kernel trick to RFA for nonlinear relational dimensionality reduction.

4.1. Global Convergence of RFA

Theorem 1 states the convergence of RFA.

Theorem 1. *The RFA algorithm converges to a global optimal solution of Problem (3).*

Proof. Considering that all the formulas, from Problem (3) to Problem (12) in the previous section, are all transformed equivalently, we only prove the convergence of Problem (12) here. Specifically, we show that $\widetilde{\mathcal{L}}$ of Problem (12) has a lower bound, which gradually decreases with the iterative process.

We can easily see that for any $\widetilde{\mathbf{W}}$, it satisfies that $0 \leq \frac{\operatorname{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}})}{\operatorname{tr}(\widetilde{\mathbf{W}}^T \widehat{\mathbf{S}}_T \widetilde{\mathbf{W}})} \leq 1$, so that the lower bound of $\widetilde{\mathcal{L}}$ is 0.

Next, we prove that the objective value of Problem (12) gradually decreases with the iterative process of RFA. Defining

$$\widetilde{\eta}_t = \frac{\mathrm{tr}(\widetilde{\mathbf{W}}_{t-1}^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}}_{t-1})}{\mathrm{tr}(\widetilde{\mathbf{W}}_{t-1}^T \widehat{\mathbf{S}}_T \widetilde{\mathbf{W}}_{t-1})}, \tag{15}$$

we then have

$$\mathrm{tr}(\widetilde{\mathbf{W}}_{t-1}^T (\widehat{\mathbf{S}}_I - \widetilde{\eta}_t \widehat{\mathbf{S}}_T) \widetilde{\mathbf{W}}_{t-1}) = 0. \tag{16}$$

However,

$$\widetilde{\mathbf{W}}_t = \underset{\widetilde{\mathbf{W}}^T \widetilde{\mathbf{W}} = \mathbf{I}_d}{\arg\min} \, \mathrm{tr}(\widetilde{\mathbf{W}}^T (\widehat{\mathbf{S}}_I - \widetilde{\eta}_t \widehat{\mathbf{S}}_T) \widetilde{\mathbf{W}}). \tag{17}$$

Therefore,

$$\mathrm{tr}(\widetilde{\mathbf{W}}_t^T (\widehat{\mathbf{S}}_I - \widetilde{\eta}_t \widehat{\mathbf{S}}_T) \widetilde{\mathbf{W}}_t) \leq 0, \tag{18}$$

and

$$\widetilde{\eta}_{t+1} = \frac{\mathrm{tr}(\widetilde{\mathbf{W}}_t^T \widehat{\mathbf{S}}_I \widetilde{\mathbf{W}}_t)}{\mathrm{tr}(\widetilde{\mathbf{W}}_t^T \widehat{\mathbf{S}}_T \widetilde{\mathbf{W}}_t)} \leq \widetilde{\eta}_t. \tag{19}$$

Thus, we prove that $\widetilde{\eta}_t$ gradually decreases with the iterative process, and Theorem 1 holds. □

According to Theorem 1, we can see that RFA can converge to the global optimal solution. In addition, for given intrinsic and penalty graphs and the relational matrix, the computational complexity of the RFA algorithm is only $\Theta(N)$, where N is the total number of data. That also illustrates the efficiency of our algorithm.

4.2. Kernel Extension

In this subsection, we apply kernel trick to our RFA method and present the kernel RFA (KRFA) method, which can be used to nonlinear dimensionality reduction problems.

The KRFA optimization problem is basically the same as Problem (3), except that the data point $\mathbf{X}_{*i}$ needs to be mapped to a reproducing kernel Hilbert space and obtain $\phi(\mathbf{X}_{*i})$, where $\phi(\cdot)$ denotes the mapping function. In addition, the corresponding intrinsic and penalty graphs and the relational matrix should also be mapped to the reproducing kernel Hilbert space. The kernel function $\mathbf{K}(\mathbf{X}_{*i}, \mathbf{X}_{*j}) = \phi(\mathbf{X}_{*i})^T \phi(\mathbf{X}_{*j})$.

We suppose the projection matrix $\mathbf{W} = \mathbf{\Phi}\mathbf{\Gamma}$. We have $\mathbf{W}^T \mathbf{\Phi} = \mathbf{\Gamma}^T \mathbf{K}$.

Normalizing and centering the data in the high-dimensional feature space, we substitute $\mathbf{K}$ with

$$\hat{\mathbf{K}} = \mathbf{K} - \mathbf{1}_N \mathbf{K} - \mathbf{K}\mathbf{1}_N + \mathbf{1}_N \mathbf{K}\mathbf{1}_N. \tag{20}$$

In this way, the learning task of RFA is able to be described as

$$\mathcal{L}_{\hat{\mathbf{K}}} = \min_{\mathbf{\Gamma}} \frac{\mathrm{tr}(\mathbf{\Gamma}^T \hat{\mathbf{K}} \widetilde{\mathbf{L}}_I \hat{\mathbf{K}}^T \mathbf{\Gamma})}{\mathrm{tr}(\mathbf{\Gamma}^T \hat{\mathbf{K}} \widetilde{\mathbf{L}}_P \hat{\mathbf{K}}^T \mathbf{\Gamma})} + \lambda \mathrm{tr}(\mathbf{\Gamma}^T \hat{\mathbf{K}} \widetilde{\mathbf{R}} \hat{\mathbf{K}}^T \mathbf{\Gamma}), \tag{21}$$

where $\widetilde{\mathbf{L}}_I$, $\widetilde{\mathbf{L}}_P$ and $\widetilde{\mathbf{R}}$ are the intrinsic graph, the penalty graph and the relational matrix, respectively. The distance of the samples in the reproducing kernel Hilbert space can be calculated by

$$D(x_i, x_j) = \sqrt{\mathbf{K}(x_i, x_i) + \mathbf{K}(x_i, x_j) - 2\mathbf{K}(x_i, x_j)}. \tag{22}$$

With reference to the derivation and transformation procedure of RFA, KRFA can be eventually transformed into

$$\widetilde{\mathcal{L}}_{\mathbf{K}} = \min_{\mathbf{V}_{\mathbf{K}}^T \mathbf{V}_{\mathbf{K}} = \mathbf{I}_d} \frac{\text{tr}(\mathbf{V}_{\mathbf{K}}^T \widetilde{\mathbf{S}}_l^{\mathbf{K}} \mathbf{V}_{\mathbf{K}})}{\text{tr}(\mathbf{V}_{\mathbf{K}}^T \widetilde{\mathbf{S}}_P^{\mathbf{K}} \mathbf{V}_{\mathbf{K}})}. \tag{23}$$

As KRFA's optimization process is similar to RFA, we still use iterative methods to solve it.

5. Experiments

In this section, extensive experiments are conducted to validate the effectiveness of our RFA and KRFA methods. For the linear case, we conduct experiments on document analysis, handwritten digits recognition, face recognition and webpage classification problems. For KRFA, we test its performance on several benchmark datasets. The results of comparative experiments are presented below.

5.1. Performance of RFA

To evaluate the performance of RFA, we selected several related dimensionality reduction methods for comparison with RFA. These methods were LDA, MFA, RDA, and PRPCA, respectively. Among them, LDA, MFA and RDA are supervised methods, while PRPCA is unsupervised. We note that, due to the flexibility in the design of the relational matrix, RFA could be either global or local. We describe this detail in the following part.

We followed MFA to construct the intrinsic and penalty graphs [14]. The numbers of nearest neighbors for constructing the intrinsic graph (k_i) and the penalty graph (k_p) were set to 5 and 20, respectively, for all the datasets.

We applied RFA to document understanding, face recognition and several other recognition tasks. In order to test the performance of RFA on document recognition tasks, we used the Ibn Sina ancient Arabic document dataset [49], USPS handwritten digits dataset (http://www.cs.nyu.edu/~roweis/data.html, accessed on 19 October 2023), two handwritten digits datasets (Optdigits and Pendigits) and one English letter dataset (Letter) from the UCI machine learning repository [50]. For face recognition tasks, we used two face datasets [51–54], the CMU PIE (http://www.face-rec.org/databases/, accessed on 19 October 2023) and the YaleB (http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html, accessed on 19 October 2023) datasets. At the same time, we used some UCI datasets, including Shuttle, Thyroid, Vowel and Waveform21, to evaluate RFA.

In our experiments, we used the classical graph Laplacian matrix, $\mathbf{L} = \mathbf{D} - \mathbf{M}$, to define the relational matrix, whose weight matrix $\mathbf{M}$ is shown as below:

$$\mathbf{M}_{ij} = \begin{cases} 1, & \text{if } \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_j) \text{ or } \mathbf{x}_j \in \mathcal{N}_k(\mathbf{x}_i), \\ 0, & \text{otherwise,} \end{cases} \tag{24}$$

where $\mathcal{N}_k(\mathbf{x}_i)$ is the set consisting of the k nearest neighborForof $\mathbf{x}_i$. For each dataset, the value of k was selected based on 5-fold cross-validation. Moreover, $\mathbf{D}$ is the diagonal degree matrix with $\mathbf{D}_{ii} = \sum_j \mathbf{M}_{ij}$.

We note that the Laplacian weight matrix formed by the above formula contains the relationship information between the sample and a certain number of its neighbors, which allows the iintegration of the local relationships between data into the supervised representation learning algorithm. At the same time, as a general model, the relationship matrix $\mathbf{R}$ in RFA can be of various forms. For example, $\mathbf{R}$ can be the centralization matrix $\mathbf{H} = \mathbf{I}_N - \frac{1}{N}\mathbf{1}\mathbf{1}^T$, where N denotes the data size and $\mathbf{1}$ is a column vector of length N with all ones.

For the PRPCA algorithm, we performed this experiment using the codes provided by the authors. For the RDA algorithm, we randomly selected the prototypes [24]. For algorithms other than RDA, we used the 1-nearest neighbor classifier to evaluate the classification performance of them.

5.1.1. Comparison on Hand-Writing Datasets

We first tested the performance of RFA on the hand-writing datasets. For clarity, the details of the used datasets are shown in Table 1.

Table 1. Statistics of the used datasets.

Dataset	Classes	Samples	Dimensions
Ibn Sina	174	20,688	200
USPS	10	9298	256
Pendigits	10	5620	16
Optdigits	10	10,992	64
Letter	26	20,000	16

Ibn Sina dataset: This dataset [55] is an ancient manuscript dataset, one image of which is shown in Figure 1, and we tried to identify the Arabic subwords on this dataset. In the experiment, we used a 50-page manuscript as the training set and 10 pages as the test set. We extracted the square-root velocity (SRV) representation [56] of the Arabic subwords. Then, we removed the outlier classes, including the classes that had less than 10 samples. Finally, we obtained a 174-class Arabic subword dataset with 17,543 samples for training and 3125 samples for test.
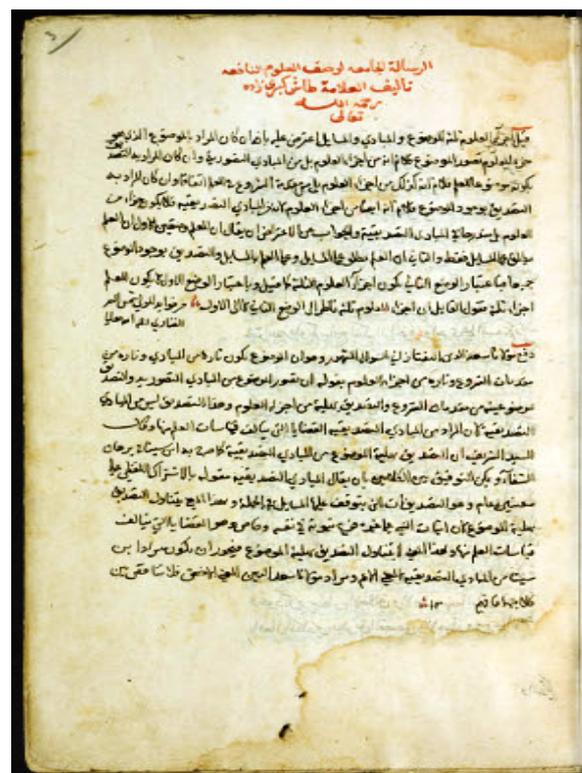

Figure 1. One image of the Ibn Sina dataset.

In this experiment, we set the number of the nearest neighbors $-k$ of the RFA to 8 and compared RFA with the LDA and MFA dimensionality reduction algorithms. The classification accuracies after using these three dimensionality reduction methods to map the data to different dimensionalities are shown in Figure 2. We can see that RFA is far better than the LDA algorithm and slightly better than the MFA algorithm. At the same time, when the dimensionality is from 50 to $C-1$ (C is the number of classes), the correct rate of the MFA algorithm has a certain fluctuation and tends to decline, while the classification performance of our RFA algorithm is relatively stable, which means that our algorithm is more robust than MFA.
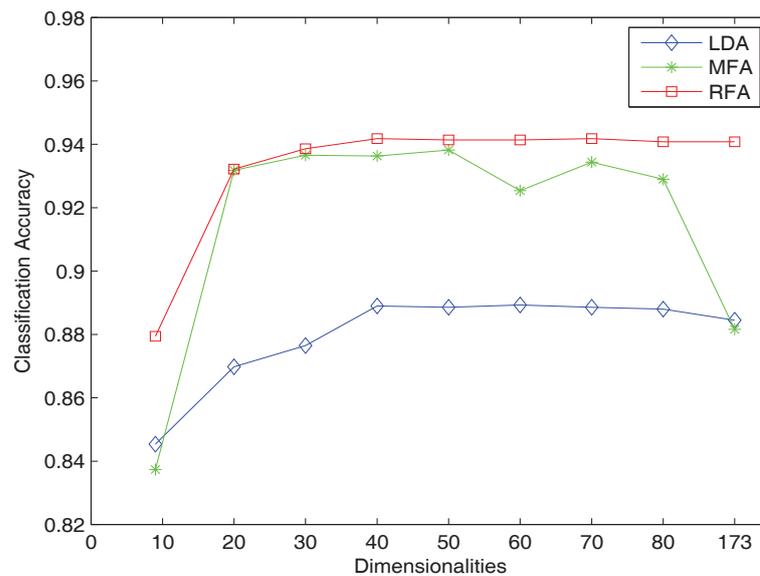
Figure 2. Classification accuracy obtained by LDA, MFA and RFA on the Ibn Sina dataset.

USPS dataset: The USPS is an U.S. post handwritten digits dataset that contains 7291 training data and 2007 test data from 10 classes and the dimensionality of the data features is 256. Some handwritten digits in the USPS dataset are shown in Figure 3.

In this experiment, we set the number of the nearest neighbors – k of RFA to 24. The classification accuracies obtained by RFA and the compared algorithms are shown in Figure 4. Because the LDA subspace has a maximum of $C-1$ dimension (C is the number of classes), LDA is only presented with a black star in the figure. We can see that when the dimension is 9, RFA obtains a comparative results with LDA and MFA. When the dimension increases, the results of RFA are always optimal. At the same time, the classification accuracy obtained by RDA is low.

Figure 5 presents 3D visualization of the learned data representations by RFA, which shows the effect of RFA to obtain better classification boundaries between the classes. In Figure 6, we show the 2D projections of data learned by both RFA and MFA, to further show the effectiveness of RFA. It is easy to see that the samples processed by RFA are less likely to overlap at the boundary, indicating that compared to MFA, RFA preserves more properties that help distinguish the samples.

In addition, the robustness of RFA is tested with respect to k_i and k_p (used to construct the intrinsic and penalty graphs). From Figure 4, it can be seen that RFA obtained the best result when the subspace dimension was 35, with parameter settings $k_i = 5$, $k_p = 20$ and $k = 24$. We fixed k and one of k_i and k_p to obtain the results when another parameter took different values. Figures 7 and 8 show that RFA is very robust.

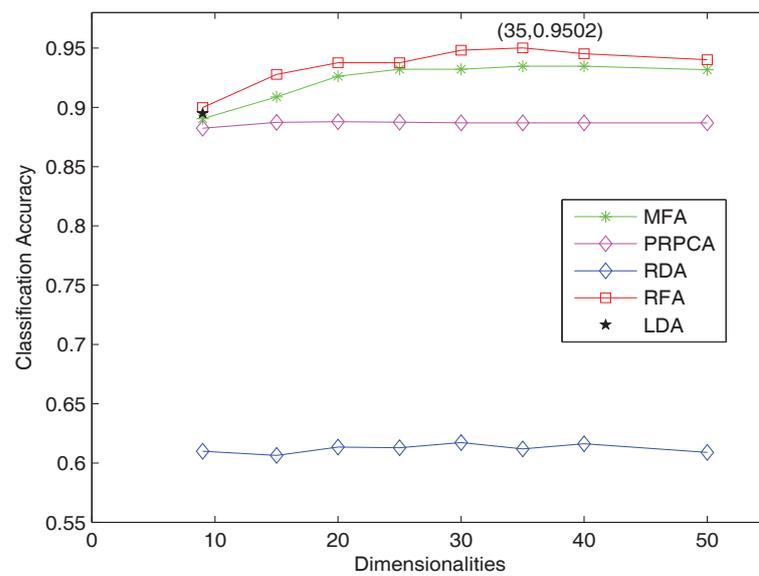**Figure 3.** Sample images from the USPS dataset.

Figure 4. Classification results obtained by RFA and the compared methods on the USPS dataset.
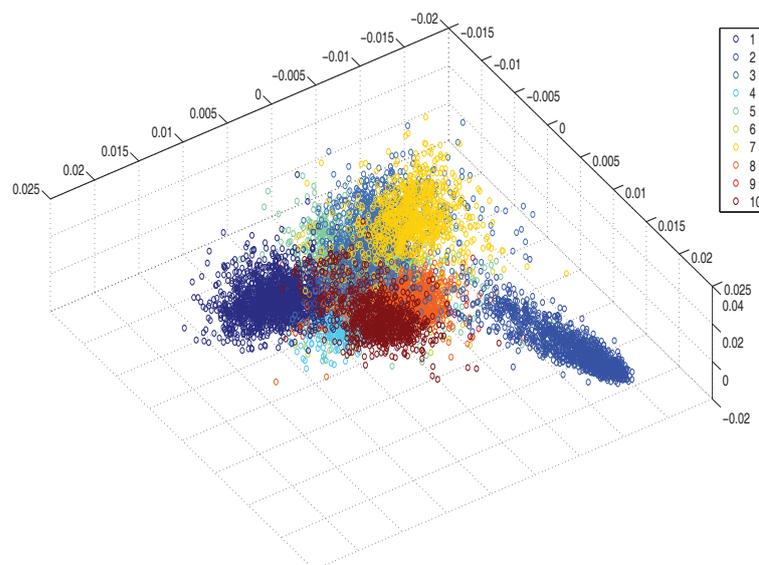

Figure 5. 3D visualization of the mapped data obtained by RFA. Samples of different classes are marked with different colors.

We also selected three document recognition-related datasets from the UCI machine learning repository to further test the effectiveness of RFA. They are Optdigits, Pendigits and Letter. Optdigits was preprocessed by NIST [57] programs to obtain 5620 instances in 8×8 dimensions. The Pendigits dataset contains a large number of preprocessed 16-dimensional samples written by 44 different authors, including 7494 training examples and 3498 test samples. Letter consists of 20,000 handwritten characters written by 20 fonts from 26 capital letters in the English alphabet. We used the 5-fold cross-validation for these experiments. Tables 2–4 show the classification accuracy and standard deviation obtained on these three datasets and the boldface results are the best ones. We can see that RFA performs consistently better than other compared methods.

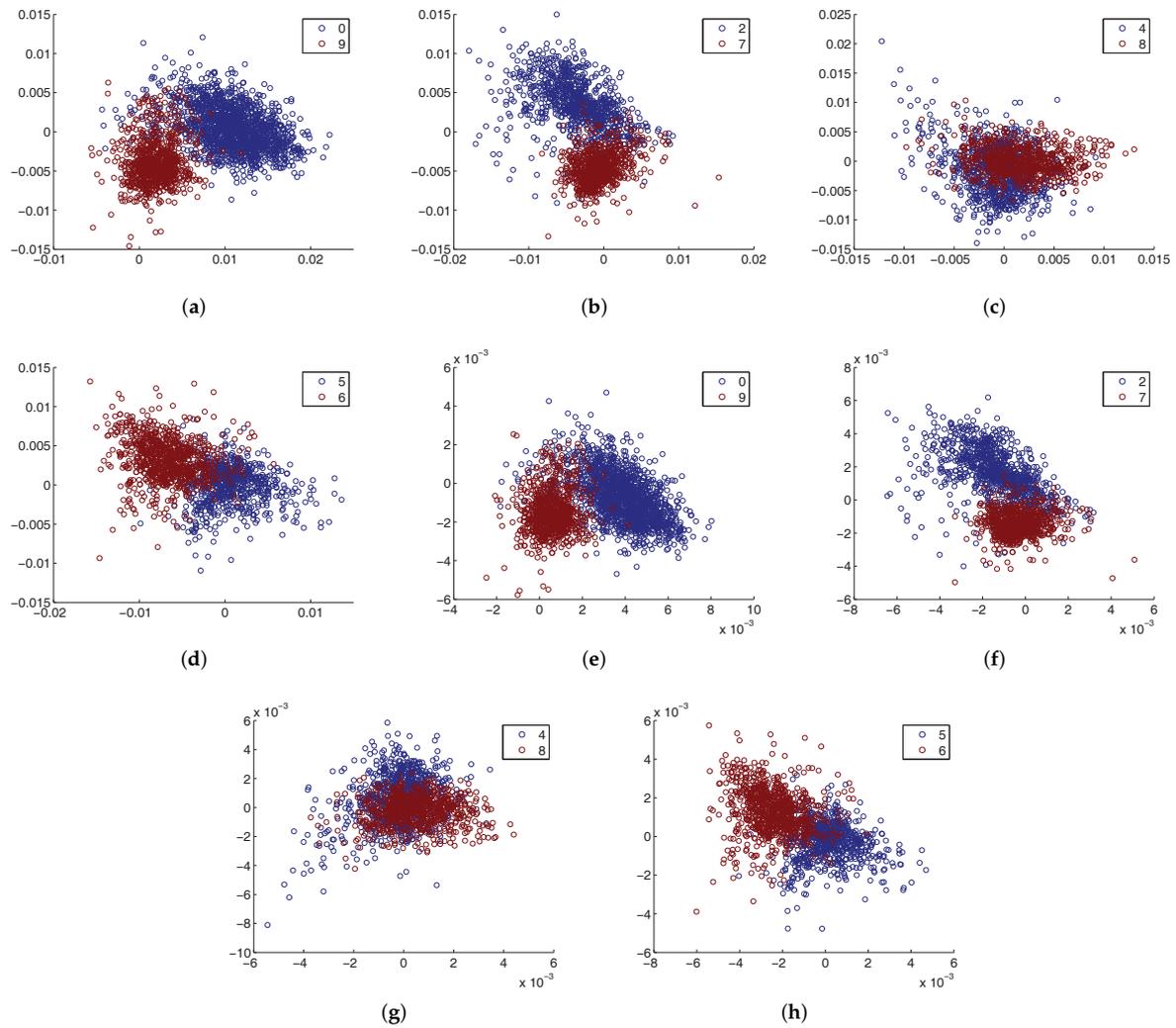**Figure 6.** 2D visualization for data pairs. (**a**–**d**) are the results obtained by RFA, and (**e**–**h**) are the results obtained by MFA.
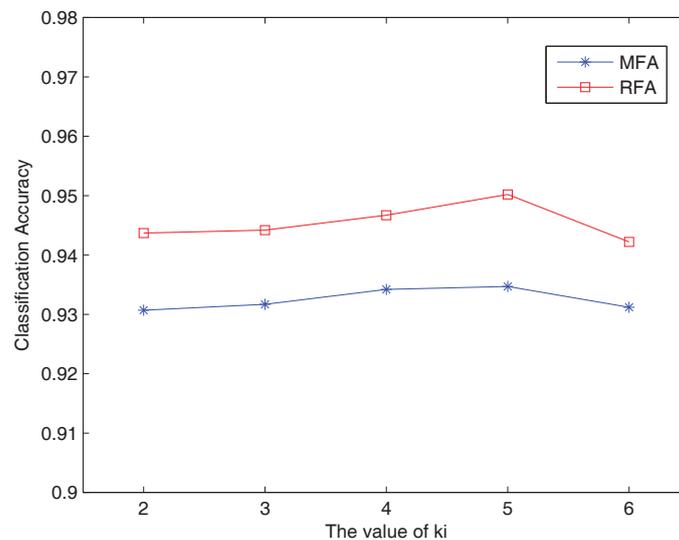

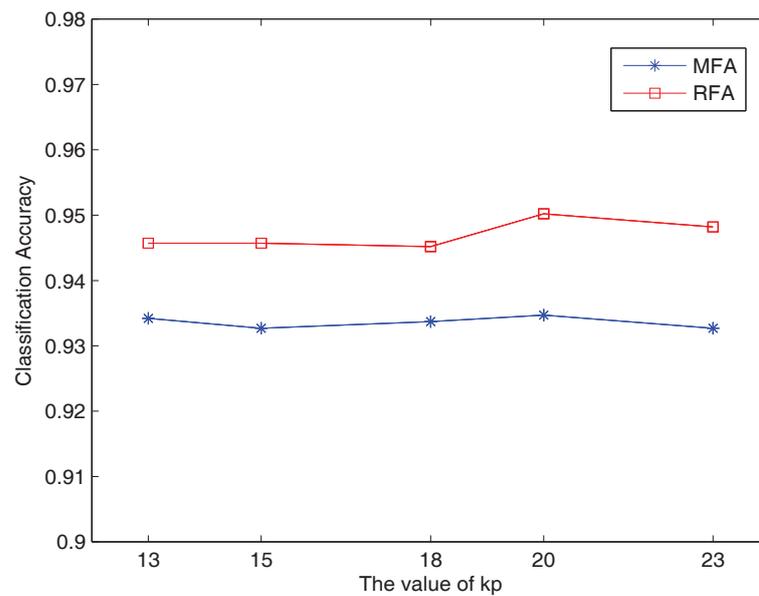**Figure 7.** Classification results obtained by RFA and MFA with different values of k_i on the USPS dataset.

Figure 8. Classification results obtained by RFA and MFA with different values of k_p on the USPS dataset.

Table 2. Classification results obtained on the Optdigits dataset. The best results are highlighted in boldface.

Dimension	LDA	MFA	PRPCA	RDA	RFA
9	0.9219 ± 0.0117	0.9657 ± 0.0020	0.9687 ± 0.0053	0.7407 ± 0.0048	$\mathbf{0.9662 \pm 0.0017}$
15	-	0.9845 ± 0.0039	0.9690 ± 0.0087	0.7407 ± 0.0017	$\mathbf{0.9875 \pm 0.0031}$
20	-	0.9836 ± 0.0041	0.9696 ± 0.0075	0.7399 ± 0.0054	$\mathbf{0.9891 \pm 0.0025}$
25	-	0.9849 ± 0.0030	0.9696 ± 0.0071	0.7425 ± 0.0042	$\mathbf{0.9891 \pm 0.0035}$
30	-	0.9819 ± 0.0062	0.9698 ± 0.0071	0.7415 ± 0.0053	$\mathbf{0.9875 \pm 0.0048}$
35	-	0.9795 ± 0.0051	0.9703 ± 0.0067	0.7404 ± 0.0066	$\mathbf{0.9879 \pm 0.0036}$
40	-	0.9783 ± 0.0058	0.9710 ± 0.0079	0.7402 ± 0.0056	$\mathbf{0.9875 \pm 0.0032}$
50	-	0.9781 ± 0.0026	0.9701 ± 0.0071	0.7407 ± 0.0043	$\mathbf{0.9877 \pm 0.0023}$

Table 3. Classification results obtained on the Pendigits dataset.

Dimension	LDA	MFA	PRPCA	RDA	RFA
9	0.9843 ± 0.0020	0.9929 ± 0.0013	0.9797 ± 0.0027	0.9203 ± 0.0046	$\mathbf{0.9939 \pm 0.0006}$
13	-	0.9940 ± 0.0026	0.9604 ± 0.0028	0.8856 ± 0.0052	$\mathbf{0.9950 \pm 0.0018}$
15	-	0.9945 ± 0.0018	0.9887 ± 0.0008	0.9212 ± 0.0050	$\mathbf{0.9962 \pm 0.0013}$

Table 4. Classification results obtained on the Letter dataset.

Dimension	LDA	MFA	PRPCA	RDA	RFA
9	0.9108 ± 0.0070	0.9570 ± 0.0014	0.9230 ± 0.0018	0.3699 ± 0.0064	$\mathbf{0.9580 \pm 0.0018}$
13	0.9108 ± 0.0070	0.9725 ± 0.0022	0.9604 ± 0.0028	0.3727 ± 0.0065	$\mathbf{0.9753 \pm 0.0022}$
15	0.9575 ± 0.0023	0.9596 ± 0.0018	0.9567 ± 0.0022	0.3715 ± 0.0087	$\mathbf{0.9695 \pm 0.0011}$

5.1.2. Comparison on Face Datasets

Here, we tested RFA on the face recognition problems. The PIE and YaleB datasets were used. The details of these two datasets are shown in Table 5. For the corresponding experimental settings, we set the number of the nearest neighbors – k on the PIE dataset to 8 and k on the YaleB dataset to 18.

Considering that RDA cannot perform well on multi-class classification problems, we used LPP instead as a compared method in these experiments. We used the 5-fold cross-validation to decide the value of parameter *k* for graph construction in LPP. We performed the experiments in different low-dimensional spaces on the PIE and YaleB datasets, and the experimental results are shown in Tables 6 and 7.

Table 5. Statistics of the face recognition datasets.

Dataset	Classes	Samples	Feature Dimension
PIE	68	11,554	1024
YaleB	38	2414	1024

Table 6. Classification results obtained on the PIE dataset.

Dimension	LDA	MFA	PRPCA	LPP	RFA
9	0.9021 ± 0.0070	0.9173 ± 0.0045	0.7736 ± 0.0113	0.8156 ± 0.0089	$\mathbf{0.9198 \pm 0.0026}$
15	0.9430 ± 0.0038	0.9526 ± 0.0040	0.8810 ± 0.0063	0.8930 ± 0.0054	$\mathbf{0.9540 \pm 0.0021}$
20	0.9528 ± 0.0031	0.9592 ± 0.0027	0.9089 ± 0.0062	0.9122 ± 0.0044	$\mathbf{0.9603 \pm 0.0030}$
25	0.9597 ± 0.0040	0.9617 ± 0.0028	0.9224 ± 0.0056	0.9262 ± 0.0042	$\mathbf{0.9621 \pm 0.0040}$
30	0.9613 ± 0.0037	0.9639 ± 0.0042	0.9299 ± 0.0041	0.9320 ± 0.0031	$\mathbf{0.9644 \pm 0.0027}$
35	0.9623 ± 0.0025	0.9643 ± 0.0024	0.9342 ± 0.0037	0.9352 ± 0.0040	$\mathbf{0.9660 \pm 0.0024}$
40	0.9633 ± 0.0019	0.9638 ± 0.0021	0.9370 ± 0.0049	0.9383 ± 0.0036	$\mathbf{0.9655 \pm 0.0021}$
50	0.9639 ± 0.0012	0.9643 ± 0.0029	0.9394 ± 0.0052	0.9411 ± 0.0034	$\mathbf{0.9656 \pm 0.0024}$
C-1(67)	0.9641 ± 0.0048	0.9640 ± 0.0034	0.9356 ± 0.0045	0.9437 ± 0.0035	$\mathbf{0.9654 \pm 0.0037}$

Table 7. Classification results obtained on the YaleB dataset.

Dimension	LDA	MFA	PRPCA	LPP	RFA
9	0.7788 ± 0.0199	$\mathbf{0.8355 \pm 0.0190}$	0.3675 ± 0.0378	0.4859 ± 0.0437	0.8347 ± 0.0212
15	0.8774 ± 0.0092	0.9031 ± 0.0071	0.4793 ± 0.0572	0.5911 ± 0.0367	$\mathbf{0.9039 \pm 0.0096}$
20	0.9209 ± 0.0136	0.9273 ± 0.0104	0.5460 ± 0.0426	0.6276 ± 0.0171	$\mathbf{0.9279 \pm 0.0067}$
25	$\mathbf{0.9395 \pm 0.0189}$	0.9370 ± 0.0126	0.5427 ± 0.0377	0.6392 ± 0.0345	0.9387 ± 0.0109
30	$\mathbf{0.9507 \pm 0.0157}$	0.9490 ± 0.0081	0.5518 ± 0.0399	0.6736 ± 0.0219	0.9495 ± 0.0079
35	$\mathbf{0.9582 \pm 0.0132}$	0.9503 ± 0.0093	0.5481 ± 0.0385	0.6740 ± 0.0109	0.9503 ± 0.0116
40	-	0.9511 ± 0.0102	0.2084 ± 0.3698	0.6997 ± 0.0059	$\mathbf{0.9569 \pm 0.0146}$
50	-	0.9482 ± 0.0101	0.3698 ± 0.1481	0.7142 ± 0.0123	$\mathbf{0.9532 \pm 0.0089}$
C-1(37)	$\mathbf{0.9594 \pm 0.0094}$	0.9511 ± 0.0122	0.5306 ± 0.2147	0.6864 ± 0.0102	0.9548 ± 0.0107

We can see from the experimental results that RFA performs very well. Although LPP is an effective dimensionality reduction method for face recognition, RFA is significantly better than LPP. Moreover, RFA obtains comparable results with LDA and MFA. These results demonstrate the effectiveness of RFA in the face recognition applications.

Additionally, convergence of RFA is verified on these two datasets. As illustrated in Figures 9 and 10, the value of η (trace ratio) decreases through the iterative procedures until it reaches the global optimal value η^* on both of the two datasets, which clearly shows the convergence of RFA.
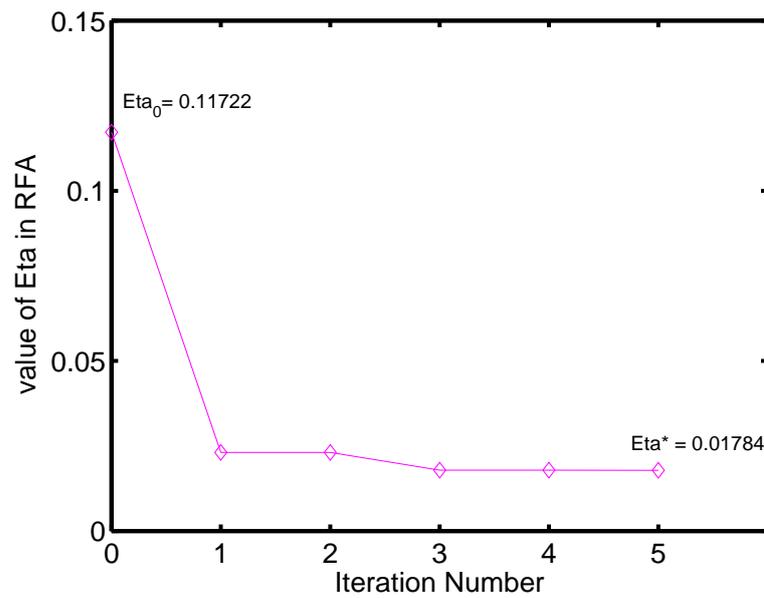
Figure 9. Changing cave of η over the iteration number on the PIE dataset.
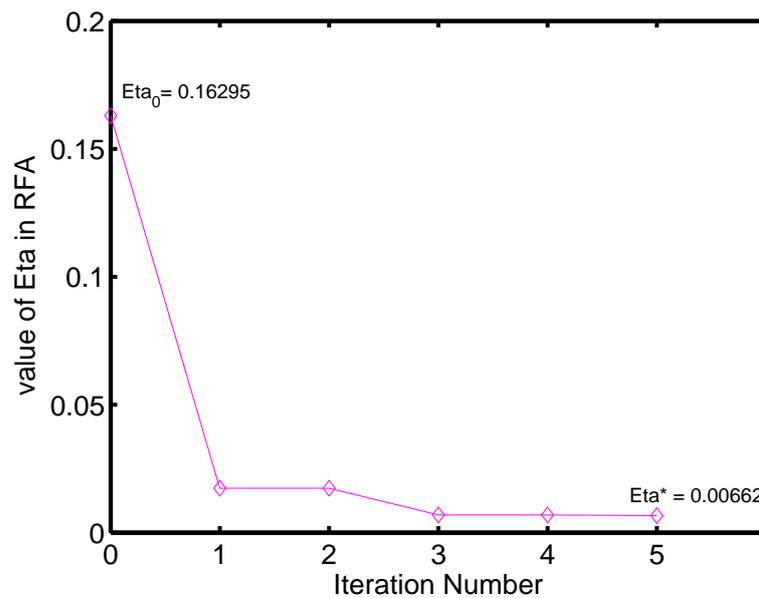

Figure 10. Changing cave of η over the iteration number on the YaleB dataset.

5.1.3. Comparison on Other UCI Datasets

To evaluate the generalization ability of RFA, we conducted experiments on UCI datasets of other fields. The details of the used datasets are shown in Table 8. For the corresponding experimental settings of these four dataset, we set the number of the nearest neighbors $-k$ to 15. For the fairness of comparison, the subspace dimension of each method was set to $C-1$.

Table 8. Statistics of the UCI datasets.

Dataset	Classes	Samples	Dimensions
Shuttle	7	14,516	9
Thyroid	3	215	5
Vowel	11	990	10
Waveform21	3	5000	21

The results shown in Table 9 demonstrate the advantage of RFA over the related approaches. It is very effective in a wide range of applications.

Table 9. Classification results obtained on several UCI datasets.

Dimension	LDA	MFA	PRPCA	RDA	RFA
Shuttle	0.9975 ± 0.0009	0.9979 ± 0.0012	0.9959 ± 0.0037	0.9162 ± 0.0132	**0.9981 ± 0.0003**
Thyroid	0.9395 ± 0.0265	0.9488 ± 0.0303	0.9488 ± 0.0195	0.9209 ± 0.0265	**0.9581 ± 0.0255**
Vowel	0.9859 ± 0.0109	0.9869 ± 0.0085	0.9859 ± 0.0090	0.3505 ± 0.0349	**0.9879 ± 0.0077**
Waveform21	0.8184 ± 0.0119	0.8258 ± 0.0197	0.8150 ± 0.0118	0.5972 ± 0.0279	**0.8278 ± 0.0175**

5.1.4. Comparison on Document Classification and Webpage Classification Problems

As a general dimensionality reduction framework, the relational matrix can be constructed with different strategies. In the previous sections, we considered the relationship between samples based on their class labels or similarity. However, in some complicated problems, relationships may presented in other forms. For example, as indicated in [25], if there is a reference relationship between two papers, they are likely to have the same topic. However, due to the sparse nature of the bag-of-words representation, the similarity between these two papers may be very low. Thus, to further testify RFA, we designed a relational matrix based on the citation relevance between data samples to model RFA, and tested its effectiveness on document classification and webpage classification problems.

For this experiment, we used two datasets, citeseer and WebKB (https://linqs-data.soe.ucsc.edu/public/lbc/, accessed on 19 October 2023). We note that the WebKB dataset contains four subsets: Cornell, Texas, Washington and Wisconsin, and we show the experimental results of these four subsets separately. Each dataset contains bag-of-words representation of documents or webpages and citation links between the instances. Citeseer contains 3312 scientific documents from 6 different classes, and there are 4732 citation relation between the documents. WebKB consists of 877 webpages from 5 different classes, and there are 1608 page links within this dataset. We adopted the same strategy as in PRPCA to construct the relational matrix:

(1) Constructing the adjacent graph $\mathbf{A}$ according to the relevance between data samples. If there was a citation or link between sample i and j, then $\mathbf{A}_{ij} = 1$; else, $\mathbf{A}_{ij} = 0$.

(2) Letting $\tilde{\mathbf{D}}_{ii} = \sum_j \mathbf{A}_{ij} = (\mathbf{A}\mathbf{A})_{ij}$, then $\mathbf{B} = \mathbf{A}\mathbf{A} - \tilde{\mathbf{D}}$,

$$\mathbf{B}_{ij} = \begin{cases} (\mathbf{A}\mathbf{A})_{ij} = \sum_{k=1}^{N} \mathbf{A}_{ij}\mathbf{A}_{kj}, & \text{if } i \neq j, \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

(3) Defining $\mathbf{G} = 2\mathbf{A} + \mathbf{B}$ as the relational matrix in RFA.

We took PRPCA as the baseline method in this part. Experimental results are illustrated in Figure 11; we can see that RFA achieves comparable results with PRPCA in all these five datasets and is even better than PRPCA on some of the datasets.
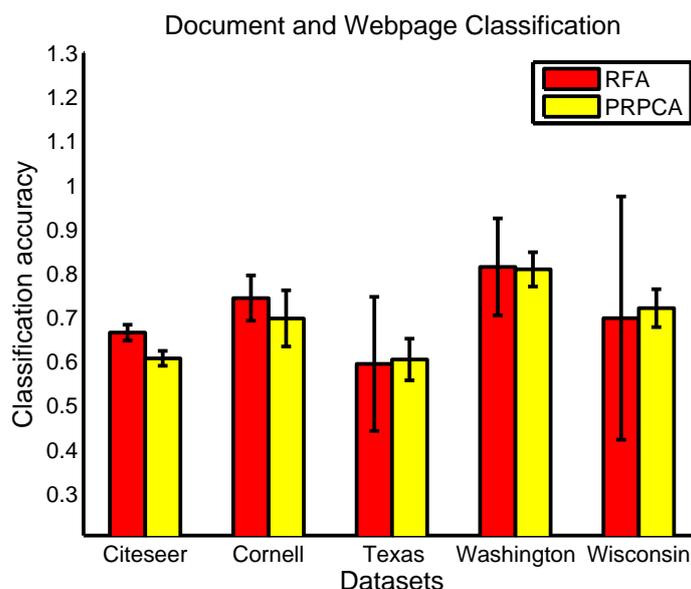
Figure 11. Classification accuracy and standard deviation obtained by RFA and PRPCA on the document and webpage classification problems.

5.2. Performance of KRFA

To evaluate the efficiency of KRFA, we tested its performance on several benchmark datasets from the UCI machine learning repository. The details of these datasets are shown in Table 10. For the corresponding experimental settings, we set the number of the nearest neighbors – k of these five dataset to 15. To avoid the singular value issue, we adopted KPCA to retain 98% of the variance before formally performing KMFA and KRFA. We used Gaussian kernel in the experiment and for the fairness of the comparison, the subspace dimension of each method was set to $C - 1$. Table 11 shows the comparison results obtained by KRFA, KMFA and RFA.

Table 10. Statistics of the UCI datasets.

Dataset	Classes	Samples	Dimensions
Ecoli	8	336	8
Satimage	7	6435	36
Vehicle	4	846	18
Waveform40	3	5000	40
Wine	3	178	13

Table 11. Classification accuracy and standard deviation obtained on several UCI datasets.

Dimension	KMFA	RFA	KRFA
Ecoli	0.4200 ± 0.0431	0.4230 ± 0.0423	**0.4511 ± 0.0648**
Satimage	0.7749 ± 0.0345	0.3915 ± 0.0834	**0.8182 ± 0.0343**
Vehicle	0.7694 ± 0.0152	0.6808 ± 0.0489	**0.7824 ± 0.0187**
Waveform40	0.7996 ± 0.0136	0.8216 ± 0.0097	**0.8256 ± 0.0062**
Wine	0.9716 ± 0.0350	0.9716 ± 0.0286	**0.9773 ± 0.0239**

As shown in Table 11, the proposed KRFA obtained a comparable and even better result than KMFA. Furthermore, experimental results of KRFA were all better than RFA on the used datasets. That superiority can be especially reflected on the Satimage dataset. The performance of RFA on Satimage was unsatisfactory. However, KRFA conducted effective nonlinear dimensionality reduction and thus obtained good result on the following

classification problem. These two points clearly demonstrate the nonlinear dimensionality reduction ability of KRFA.

6. Conclusions

In this paper, we propose a novel and general framework named relational Fisher analysis (RFA) which integrates relational information into the dimensionality reduction models. RFA can be effectively optimized with an iterative method based on trace ratio. For nonlinear dimensionality reduction, we adopt kernel trick to RFA and design its kernel version named KRFA. Extensive experiments demonstrate that RFA and KRFA outperform other related dimensionality reduction algorithms in most cases. In future work, we plan to extend this research in the following aspects: (1) Exploiting efficient relationship metric for different relational data to further test the effectiveness of the proposed RFA model; (2) Further extending the formulation of RFA for semi-supervised learning; and (3) Extending RFA for tensor representation learning and applying it to tensor analysis problems.

Author Contributions: Conceptualization, L.-N.W. and G.Z.; methodology, L.-N.W. and M.C.; software, Y.S.; validation, L.-N.W. and M.C.; formal analysis, G.Z.; investigation, L.-N.W.; resources, G.Z.; data curation, Y.S.; writing—original draft preparation, L.-N.W., Y.S. and G.Z.; writing—review and editing, M.C.; visualization, Y.S.; supervision, G.Z.; project administration, M.C.; funding acquisition, G.Z. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used in this work were public on the Internet and no new data were created.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jolliffe, I. *Principal Component Analysis*; Wiley Online Library: Hoboken, NJ, USA, 2002.
2. Martínez, A.M.; Kak, A.C. Pca versus lda. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 228–233. [CrossRef]
3. Turk, M.A.; Pentland, A.P. Face recognition using eigenfaces. In Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Maui, HI, USA, 3–6 June 1991; pp. 586–591.
4. Fukunaga, K. *Introduction to Statistical Pattern Recognition*; Academic Press: Cambridge, MA, USA, 2013.
5. Ye, J.; Janardan, R.; Park, C.H.; Park, H. An optimization criterion for generalized discriminant analysis on undersampled problems. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 982–994. [PubMed]
6. Lu, J.; Plataniotis, K.N.; Venetsanopoulos, A.N. Face recognition using LDA-based algorithms. *IEEE Trans. Neural Netw.* **2003**, *14*, 195–200. [PubMed]
7. Fisher, R.A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* **1936**, *7*, 179–188. [CrossRef]
8. Schölkopf, B.; Smola, A.J.; Müller, K. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Comput.* **1998**, *10*, 1299–1319. [CrossRef]
9. Baudat, G.; Anouar, F. Generalized Discriminant Analysis Using a Kernel Approach. *Neural Comput.* **2000**, *12*, 2385–2404. [CrossRef]
10. Zhao, H.; Wang, Z.; Nie, F. A New Formulation of Linear Discriminant Analysis for Robust Dimensionality Reduction. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 629–640. [CrossRef]
11. Tenenbaum, J.B.; de Silva, V.; Langford, J.C. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* **2000**, *290*, 2319–2323. [CrossRef]

12.   Roweis, S.T.; Saul, L.K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323–2326. [CrossRef]
13.   Belkin, M.; Niyogi, P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Comput.* **2003**, *15*, 1373–1396. [CrossRef]
14.   Yan, S.; Xu, D.; Zhang, B.; Zhang, H.J.; Yang, Q.; Lin, S. Graph Embedding and Extensions: A General Framework for Dimensionality Reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 40–51. [CrossRef] [PubMed]
15.   He, X.; Niyogi, P. Locality Preserving Projections. In Proceedings of the 16th International Conference on Neural Information Processing Systems, Whistler, BC, Canada, 9–11 December 2003; pp. 153–160.
16.   Wong, W.K.; Zhao, H. Supervised optimal locality preserving projection. *Pattern Recognit.* **2012**, *45*, 186–197. [CrossRef]
17.   Chen, J.; Liu, Y. Locally linear embedding: A survey. *Artif. Intell. Rev.* **2011**, *36*, 29–48. [CrossRef]
18.   Wang, Q.; Chen, K. Zero-Shot Visual Recognition via Bidirectional Latent Embedding. *Int. J. Comput. Vis.* **2017**, *124*, 356–383. [CrossRef]
19.   Chang, J.; Blei, D.M. Relational Topic Models for Document Networks. *J. Mach. Learn. Res. Proc. Track* **2009**, *5*, 81–88.
20.   Chang, J.; Boyd-Graber, J.L.; Blei, D.M. Connections between the Lines: Augmenting Social Networks with Text. In Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009; pp. 169–178.
21.   Peel, L. Graph-based semi-supervised learning for relational networks. *arXiv* **2016**, arXiv:1612.05001.
22.   Yang, Z.; Cohen, W.; Salakhutdinov, R. Revisiting semi-supervised learning with graph embeddings. *arXiv* **2016**, arXiv:1603.08861.
23.   Weston, J.; Ratle, F.; Mobahi, H.; Collobert, R. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 639–655.
24.   Duin, R.P.W.; Pekalska, E.; de Ridder, D. Relational Discriminant Analysis. *Pattern Recognit. Lett.* **1999**, *20*, 1175–1181. [CrossRef]
25.   Li, W.J.; Yeung, D.Y.; Zhang, Z. Probabilistic Relational PCA. In Proceedings of the 22nd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 7–10 December 2009; pp. 1123–1131.
26.   Gao, K.; Liu, B.; Yu, X.; Qin, J.; Zhang, P.; Tan, X. Deep Relation Network for Hyperspectral Image Few-Shot Classification. *Remote Sens.* **2020**, *12*, 923. [CrossRef]
27.   Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Li, J.; Ji, R. Local Relation Learning for Face Forgery Detection. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, Virtually, 2–9 February 2021; pp. 1081–1088.
28.   Cho, M.; Kim, M.; Hwang, S.; Park, C.; Lee, K.; Lee, S. Look Around for Anomalies: Weakly-Supervised Anomaly Detection via Context-Motion Relational Learning. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 17–24 June 2023; pp. 12137–12146.
29.   Ma, R.; Mei, B.; Ma, Y.; Zhang, H.; Liu, M.; Zhao, L. One-shot relational learning for extrapolation reasoning on temporal knowledge graphs. *Data Min. Knowl. Discov.* **2023**, *37*, 1591–1608. [CrossRef]
30.   Li, X.; Ma, J.; Yu, J.; Zhao, M.; Yu, M.; Liu, H.; Ding, W.; Yu, R. A structure-enhanced generative adversarial network for knowledge graph zero-shot relational learning. *Inf. Sci.* **2023**, *629*, 169–183. [CrossRef]
31.   Zhong, G.; Shi, Y.; Cheriet, M. Relational Fisher analysis: A general framework for dimensionality reduction. In Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, 24–29 July 2016; pp. 2244–2251.
32.   He, X.; Yan, S.; Hu, Y.; Zhang, H.J. Learning a Locality Preserving Subspace for Visual Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Nice, France, 13–16 October 2003; Volume 1, pp. 385–392.
33.   Wang, H.; Yan, S.; Xu, D.; Tang, X.; Huang, T.S. Trace Ratio vs. Ratio Trace for Dimensionality Reduction. In Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, MN, USA, 18–23 June 2007.
34.   Nie, F.; Xiang, S.; Jia, Y.; Zhang, C.; Yan, S. Trace Ratio Criterion for Feature Selection. In Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, IL, USA, 13–17 July 2008; Volume 2, pp. 671–676.
35.   Zhong, G.; Ling, X. The necessary and sufficient conditions for the existence of the optimal solution of trace ratio problems. In Proceedings of the Chinese Conference on Pattern Recognition, Chengdu, China, 5–7 November 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 742–751.
36.   Duin, R.P.; de Ridder, D.; Tax, D.M. Experiments with a featureless approach to pattern recognition. *Pattern Recognit. Lett.* **1997**, *18*, 1159–1166. [CrossRef]
37.   Duin, R.P. Relational discriminant analysis and its large sample size problem. In Proceedings of the 14th International Conference on Pattern Recognition, Washington, DC, USA, 16–20 August 1998; Volume 1, pp. 445–449.
38.   Ypma, A.; Duin, R.P. Support objects for domain approximation. In Proceedings of the 8th International Conference on Artificial Neural Networks, Skövde, Sweden, 2–4 September 1998; Springer: Berlin/Heidelberg, Germany, 1998; pp. 719–724.
39.   Borg, I.; Groenen, P.J. *Modern Multidimensional Scaling: Theory and Applications*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2005.
40.   Sammon, J., Jr. A nonlinear structure analysis mapping for data. *IEEE Trans. Comp.* **1969**, *C-18*, 401–409.
41.   Niemann, H. Linear and nonlinear mapping of patterns. *Pattern Recognit.* **1980**, *12*, 83–87. [CrossRef]
42.   Paccanaro, A.; Hinton, G.E. Learning Distributed Representations of Concepts Using Linear Relational Embedding. *IEEE Trans. Knowl. Data Eng.* **2001**, *13*, 232–244. [CrossRef]

43. Wang, H.; Li, W. Relational Collaborative Topic Regression for Recommender Systems. *IEEE Trans. Knowl. Data Eng.* **2015**, *27*, 1343–1355. [CrossRef]

44. Xuan, J.; Lu, J.; Zhang, G.; Xu, R.Y.D.; Luo, X. Bayesian Nonparametric Relational Topic Model through Dependent Gamma Processes. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 1357–1369. [CrossRef]

45. Jia, Y.; Nie, F.; Zhang, C. Trace Ratio Problem Revisited. *IEEE Trans. Neural Networks* **2009**, *20*, 729–735.

46. Xiang, S.; Nie, F.; Zhang, C. Learning A Mahalanobis Distance Metric for Data Clustering and Classification. *Pattern Recognit.* **2008**, *41*, 3600–3612. [CrossRef]

47. Nie, F.; Xiang, S.; Jia, Y.; Zhang, C. Semi-supervised Orthogonal Discriminant Analysis via Label Propagation. *Pattern Recognit.* **2009**, *42*, 2615–2627. [CrossRef]

48. Zhao, M.B.; Zhang, Z.; Chow, T.W.S. Trace Ratio Criterion Based Generalized Discriminative Learning for Semi-supervised Dimensionality Reduction. *Pattern Recognit.* **2012**, *45*, 1482–1499. [CrossRef]

49. Moghaddam, R.F.; Cheriet, M.; Milo, T.; Wisnovsky, R. A Prototype System for Handwritten Sub-word Recognition: Toward Arabic-manuscript Transliteration. In Proceedings of the 2012 11th International Conference on Information Science, Signal Processing and Their Applications (ISSPA), Montreal, QC, Canada, 2–5 July 2012; pp. 1198–1204.

50. Lichman, M. UCI Machine Learning Repository, 2013. Available online: https://archive.ics.uci.edu/ (accessed on 19 October 2023).

51. Gross, R. Face Databases. In *Handbook of Face Recognition*; Li, S.Z., Jain, A.K., Eds.; Springer: Berlin/Heidelberg, Germany, 2005.

52. Cai, D.; He, X.; Hu, Y.; Han, J.; Huang, T. Learning a Spatially Smooth Subspace for Face Recognition. In Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), Minneapolis, MN, USA, 18–23 June 2007.

53. Cai, D.; He, X.; Han, J. Spectral Regression for Efficient Regularized Subspace Learning. In Proceedings of the 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, 14–20 October 2007.

54. Cai, D.; He, X.; Han, J.; Zhang, H.J. Orthogonal Laplacianfaces for face recognition. *IEEE Trans. Image Process.* **2006**, *15*, 3608–3614. [CrossRef]

55. Zheng, Y.; Cai, Y.; Zhong, G.; Chherawala, Y.; Shi, Y.; Dong, J. Stretching Deep Architectures for Text Recognition. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 236–240.

56. Srivastava, A.; Klassen, E.; Joshi, S.H.; Jermyn, I.H. Shape Analysis of Elastic Curves in Euclidean Spaces. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1415–1428. [CrossRef]

57. Garris, M.D.; Blue, J.L.; Candela, G.T.; Dimmick, D.L.; Geist, J.; Grother, P.J.; Janet, S.A.; Wilson, C.L. *NIST Form-Based Handprint Recognition System*; NIST Interagency/Internal Report (NISTIR); National Institute of Standards and Technology: Gaithersburg, MD, USA, 1994.