machine learning & knowledge extraction

Article

Recent Advances in Supervised Dimension Reduction: A Survey

Guoqing Chao [1], **Yuan Luo** [2] and **Weiping Ding** [3,*]

[1] School of Information Systems, Singapore Management University, Singapore 178902, Singapore; guoqingchao10@gmail.com
[2] Feinberg School of Medicine, Northwestern University, Chicago, IL 60611, USA; yuan.luo@northwestern.edu
[3] School of Computer Science and Technology, Nantong University, Nantong 226019, China
[*] Correspondence: dwp9988@163.com; Tel.: +86-513-85012517

Received: 28 October 2018; Accepted: 28 December 2018; Published: 7 January 2019

Abstract: Recently, we have witnessed an explosive growth in both the quantity and dimension of data generated, which aggravates the high dimensionality challenge in tasks such as predictive modeling and decision support. Up to now, a large amount of unsupervised dimension reduction methods have been proposed and studied. However, there is no specific review focusing on the supervised dimension reduction problem. Most studies performed classification or regression after unsupervised dimension reduction methods. However, we recognize the following advantages if learning the low-dimensional representation and the classification/regression model simultaneously: high accuracy and effective representation. Considering classification or regression as being the main goal of dimension reduction, the purpose of this paper is to summarize and organize the current developments in the field into three main classes: PCA-based, Non-negative Matrix Factorization (NMF)-based, and manifold-based supervised dimension reduction methods, as well as provide elaborated discussions on their advantages and disadvantages. Moreover, we outline a dozen open problems that can be further explored to advance the development of this topic.

Keywords: supervised learning; dimension reduction; representation learning; principal component analysis; nonnegative matrix factorization; manifold learning

1. Introduction

Currently, high-dimensional data are very common in the real world. For example, with the advance of the next generation sequencing technique, millions of SNPs (Single Nucleotide Polymorphisms) can be obtained in the Human Genome Project (HGP). Another example is digital images: a 1024 × 1024 image amounts to a 1,048,576-dimensional vector when concatenating rows or columns. In fact, high dimensionality frequently appears in time series data, medical data, and sensor data. Although the data dimension is high, often, only a small amount of key factors are important for a particular modeling task. For instance, often, up to a few hundred SNPs are implicated in a certain disease phenotype, yet the majority of the millions of other SNPs have little association with that disease [1–3]. How to identify the important variables or features and help further analysis is a fundamental problem in machine learning and many other application fields. Dimension reduction is the main topic related to this problem, and it refers to the transformation of high-dimensional data to a low-dimensional representation. Feature selection and feature extraction are two popular techniques to implement dimension reduction. Feature selection aims to select an effective subset of the existing variables [4,5], while feature extraction learns a low-dimensional combination of the existing variables [6]. Feature selection is very important in some applications such as identifying a few disease-associated SNPs across the genome. The Least Absolute Shrinkage and Selection Operator (LASSO) is a typical example of a feature selection technique. Compared with feature selection, feature extraction has attracted more attention in the past several decades, and numerous branches have seen extensive development, including Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Non-negative Matrix Factorization (NMF), the Laplacian Eigenmap (LE), Locally Linear Embedding (LLE), etc.

Most of the general dimension reduction methods belong to the unsupervised learning category because no label information is used. The other two traditional machine learning categories are supervised learning and semi-supervised learning, which use all or a part of the label information. In most real applications, dimension reduction is just an intermediate step toward the final goals, like classification or regression. Separating the dimension reduction and model learning may not be optimal for classification or regression. For example, in the task of document classification, feature selection or feature extraction methods are used first to get a low-dimensional text representation, and then, a classifier is trained to make a prediction [7,8]. Lacking supervision, some important words may be filtered before training the classifier, which affects the final performance [9]. To tackle this problem, supervised dimension reduction methods have emerged and attracted growing attention.

Based on the underlying techniques adopted, we categorize the supervised dimension reduction methods into three classes: PCA-based, NMF-based, and manifold-based dimension reduction methods. Among them, most of PCA-based and NMF-based methods are linear methods, while most of manifold-based methods are non-linear methods. By analyzing the means of exploiting the label information, we find that there are two main ways: LDA and directly integrating the loss function for classification or regression. LDA minimizes the distance within class and maximizes the distance between classes. To integrate the loss function directly for classification or regression, the commonly-used loss functions (e.g., L_2 loss, L_1 loss, and hinge loss) are mainly adopted in logistic regression, Support Vector Machine (SVM), linear regression, polynomial regression, etc. We will elaborate on them in the subsequent sections.

In the past few decades, dimension reduction had been extensively explored, and several reviews [10–17] on dimension reduction already exist. However, different from those that mainly reviewed existing unsupervised dimension reduction methods, our review focuses on the supervised dimension reduction. To the best of our knowledge, this is the first review to target this direction. We provide a taxonomy to systematically categorize the methods and list important open problems to guide the further development of this topic. Due to the greater popularity of feature extraction compared with feature selection, in our paper, we mainly focus on feature extraction for supervised learning. With regard to feature selection for supervised learning, we refer the reader to [18].

In the rest of this paper, we provide a formal definition and the taxonomy of supervised dimension reduction in Section 2. In Section 3, we describe supervised dimension reduction methods and their three classes in more detail. Section 4 reviews the real-world applications in which supervised dimension reduction methods are used. In Section 5, several promising future directions that need further exploration are unfolded. Finally, we conclude in Section 6.

2. Definition and Taxonomy

Given the data matrix $\boldsymbol{X}^{N \times D}$ and label vector $\boldsymbol{Y}^{N}$, where N indicates the number of data points and D indicates the dimension of the data, general dimension reduction seeks for a representation $\boldsymbol{U}^{N \times d}$ where $d << D$, to keep as much information as possible. It is worth noting that different general dimension reduction methods retain the information under different assumptions. For example, PCA tries to keep the information by maximizing the variance, while LE aims to keep the manifold information. For supervised dimension reduction, the final result is still the low-dimensional representation $\boldsymbol{U}^{N \times d}$, but this representation will be guided to predict the label $\boldsymbol{Y}^{N}$ by using the label information during the dimension reduction process. Using the label information $\boldsymbol{Y}^{N}$ is the main difference between supervised dimension reduction and unsupervised dimension reduction methods.

To obtain a whole picture of the existing supervised dimension reduction methods, we provide Figure 1 to show the taxonomy of supervised and semi-supervised dimension reduction techniques.

For simplicity, afterwards, we will just use supervised dimension reduction to include supervised and semi-supervised dimension reduction. We categorize the existing supervised dimension reduction methods into three classes: PCA-based, NMF-based, and manifold-based methods. For NMF-based supervised dimension reduction methods, we further divide them into two subclasses based on the way of using label information.

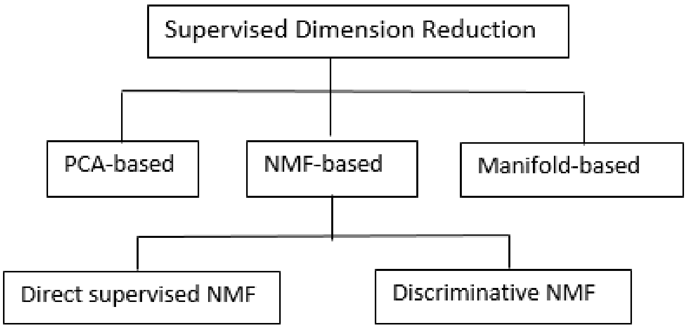

Figure 1. The taxonomy of supervised dimension reduction methods.

3. Supervised Dimension Reduction

3.1. PCA-Based Supervised Dimension Reduction

PCA can be considered as the most popular dimension reduction technique. It tries to learn the orthogonal projection of the original data onto a lower dimensional linear space, known as the principal subspace, such that the variance of the projected data is maximized [19].

To help understanding, consider the projection to a one-dimensional space ($d = 1$). For convenience, the projection vector is defined as $\boldsymbol{u}_1$ with the constraint $\boldsymbol{u}_1^T\boldsymbol{u}_1 = 1$. The mean of the projected data is $\boldsymbol{u}_1^T\bar{\boldsymbol{x}}$, where the sample mean is defined by:

$$\bar{\boldsymbol{x}} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i. \tag{1}$$

The variance of the projected data is given by $\frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{u}_1^T\boldsymbol{x}_i - \boldsymbol{u}_1^T\bar{\boldsymbol{x}})^2 = \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1$, where $\boldsymbol{S}$ is the data covariance matrix defined by:

$$\boldsymbol{S} = \frac{1}{N}\sum_{i=1}^{N}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T = \mathrm{Cov}(\boldsymbol{X}). \tag{2}$$

Now, PCA can be formulated as an optimization problem as follows:

$$\begin{cases} \max\limits_{\boldsymbol{u}_1} \boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 \\ \boldsymbol{u}_1^T\boldsymbol{u}_1 = 1. \end{cases} \tag{3}$$

By introducing Lagrange multiplier λ_1 and setting the derivative of the Lagrange function with respect to $\boldsymbol{u}_1$ equal to zero, we obtain:

$$\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1\boldsymbol{u}_1, \tag{4}$$

which shows that $\boldsymbol{u}_1$ is the eigenvector of $\boldsymbol{S}$. Left multiply the above equation by $\boldsymbol{u}_1{}^T$, and use the constraint $\boldsymbol{u}_1^T u_1 = 1$; the variance becomes:

$$\boldsymbol{u}_1^T\boldsymbol{S}\boldsymbol{u}_1 = \lambda_1. \tag{5}$$

Therefore, when $\boldsymbol{u}_1$ is set to the eigenvector corresponding to the largest eigenvalue, the variance of the data will be maximized, and this eigenvector is known as the first principal component. The subsequent principal components will be obtained by choosing a new direction that maximizes the projected variance among all possible directions orthogonal to those already considered. If d dimensional projection space is considered, the d eigenvectors $\boldsymbol{u}_1, \cdots, \boldsymbol{u}_d$ of the data covariance matrix S corresponding to the d largest eigenvalues $\lambda_1, \cdots, \lambda_d$ are the projection matrices we seek. Let $\boldsymbol{U} = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_d]$; $\boldsymbol{XU}$ will be the low-dimensional representation. Note that in PCA-related methods, $\boldsymbol{U}$ represents the projection matrix and is not the low-dimensional representation.

One heuristic to perform supervised PCA is to first select a subset of the original features based on their correlation with the label information and then apply the conventional PCA to the subset of the features to conduct dimension reduction [20]. In [21], an independence criterion named the Hilbert–Schmidt independence criterion [22] in Reproducing Kernel Hilbert Space (RKHS) is used to measure the dependence between the two variables $\mathcal{X}$ and $\mathcal{Y}$ by computing the Hilbert–Schmidt norm of the cross-covariance operator associated with their RKHSs.

Define two separable RKHSs $\mathcal{F}$ and $\mathcal{G}$ containing all continuous, bounded, and real-valued functions of x from $\mathcal{X}$ to $\mathbb{R}$ and y from $\mathcal{Y}$ to $\mathbb{R}$, respectively. Then, the cross-covariance between elements of $\mathcal{F}$ and $\mathcal{G}$ is $\mathrm{Cov}(f(x), g(y)) = \mathbf{E}_{x,y}[f(x)g(y)] - \mathbf{E}_x[f(x)]\mathbf{E}_y[g(y)]$. There is a unique linear operator $C_{x,y} : \mathcal{G} \to \mathcal{F}$ mapping elements of $\mathcal{G}$ to the elements of $\mathcal{F}$ such that $< f, C_{x,y}g >= \mathrm{Cov}(f(x), g(y)) \quad \forall f \in \mathcal{F}, \quad \forall g \in \mathcal{G}$. According to [23], this operator can be defined as $C_{x,y} = \mathbb{E}_{x,y}[(\phi(x) - \mathbb{E}_x[\phi(x)]) \otimes (\psi(y) - \mathbb{E}_y[\psi(y)])]$, where × indicates the tensor product and ϕ and ψ are the associated feature maps of $\mathcal{F}$ and $\mathcal{G}$, respectively.

Now, the Hilbert–Schmidt (HS) norm of this operator $C : \mathcal{G} \to \mathcal{F}$ is defined as $\|C\|^2_{HS} = \sum_{i,j} < Cw_i, h_j >^2_{\mathcal{F}}$ where w_i and h_j are orthogonal bases of $\mathcal{F}$ and $\mathcal{G}$, respectively. Assume $P_{\mathcal{X},\mathcal{Y}}$ is the joint distribution of variables $\mathcal{X}$ and $\mathcal{Y}$. HSIC, the square of the HS norm of the cross-covariance operator, can be expressed in terms of kernel functions as:

$$\mathrm{HSIC}(P_{\mathcal{X},\mathcal{Y}}) = \boldsymbol{E}_{x,x',y,y'}[k(x,x')l(y,y')] + \boldsymbol{E}_{x,x'}[k(x,x')]\boldsymbol{E}_{y,y'}[l(y,y')] - 2\boldsymbol{E}_{x,y}[\boldsymbol{E}'_x[k(x,x')]\boldsymbol{E}'_y[l(y,y')]], \tag{6}$$

where k and l are the associated kernel functions of $\mathcal{F}$ and $\mathcal{G}$, respectively. $\boldsymbol{E}_{x,x',y,y'}$ indicates the expectation over independent pairs of (x,y) and (x',y') drawn from $P_{\mathcal{X},\mathcal{Y}}$. In real applications, we will use an empirical estimate of HSIC. Suppose data $\mathcal{Z} = (x_1, y_1), \cdots, (x_N, y_N) \subset \mathcal{X} \times \mathcal{Y}$ are drawn independently from $P_{\mathcal{X},\mathcal{Y}}$. The empirical estimate of HSIC is given by:

$$\mathrm{HSIC}(\mathcal{Z}, \mathcal{F}, \mathcal{G}) = (N-1)^2 \boldsymbol{tr}(\boldsymbol{KHLH}), \tag{7}$$

where $\boldsymbol{H}, \boldsymbol{K}, \boldsymbol{L} \in \mathbf{R}^{\mathbf{n} \times \mathbf{n}}, K_{ij} = k(x_i, x_j), L_{ij} = l(y_i, y_j)$ and $\boldsymbol{H} = \boldsymbol{I} - N^{-1}\boldsymbol{ee}^T$ is the centering matrix.

After introducing HSIC, we will introduce the supervised PCA method using HSIC. The problem is to seek for the subspace $\boldsymbol{U}^T\boldsymbol{X}^T$ such that the dependence between the projected data $\boldsymbol{U}^T\boldsymbol{X}^T$ and the label matrix $\boldsymbol{Y}$ is maximized. It can be formulated as:

$$\begin{cases} \max\limits_{\boldsymbol{U}} \boldsymbol{tr}(\boldsymbol{KHLH}) = \boldsymbol{tr}(\boldsymbol{XUU}^T\boldsymbol{X}^T\boldsymbol{HLH}) = \boldsymbol{tr}(\boldsymbol{U}^T\boldsymbol{X}^T\boldsymbol{HLHXU}) \\ \qquad\qquad\qquad s.t. \quad \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}. \end{cases} \tag{8}$$

Obviously, this optimization problem has a closed-form solution. The eigenvectors $u_1, \cdots, u_d$ corresponding to the d largest eigenvalues $\lambda_1, \cdots, \lambda_d$ of the symmetric matrix $\boldsymbol{X}^T\boldsymbol{HLHX}$ form the optimal solution $\boldsymbol{U} = [u_1, \cdots, u_d]$. It is noted that when $\boldsymbol{L} = \boldsymbol{I}$, supervised PCA [21] degenerates to the traditional PCA.

Bin et al. [24] compared the supervised PCA with four traditional regression methods and illustrated the superiority of supervised PCA. Roberts and Martin [25] applied supervised PCA proposed in [20] to assess multiple pollutant effects. Yu et al. [26] proposed a supervised probabilistic PCA that possesses good interpretability and can handle missing values.

3.2. NMF-Based Supervised Dimension Reduction

NMF [27] aims to factorize the data matrix $\boldsymbol{X}$ into two nonnegative matrices: one is the representation (coefficient) matrix $\boldsymbol{U}^{N\times d}$ and the other one is the basis matrix $\boldsymbol{V}^{d\times D}$. The general NMF is formulated as:

$$\begin{cases} \min\limits_{\boldsymbol{U},\boldsymbol{V}} \boldsymbol{X} = \boldsymbol{UV} \\ \boldsymbol{U} \geq 0 \quad \boldsymbol{V} \geq 0. \end{cases} \tag{9}$$

NMF can be considered as approximating the true data matrix $\boldsymbol{X}$ with data matrix $\boldsymbol{Z}$, which exactly equals $\boldsymbol{UV}$. Two main loss functions are adopted to measure the divergence between $\boldsymbol{X}$ and $\boldsymbol{Z}$; one is the Frobenius loss function, and the other one is the generalized Kullback–Leibler divergence (I-divergence [28]) function. Corresponding to these two loss functions, two NMF versions are formulated as:

$$\begin{cases} \min\limits_{\boldsymbol{U},\boldsymbol{V}} \|\boldsymbol{X} - \boldsymbol{Z}\|_F^2 \\ \quad \boldsymbol{Z} = \boldsymbol{UV} \\ \boldsymbol{U} \geq 0 \quad \boldsymbol{V} \geq 0, \end{cases} \tag{10}$$

and:

$$\begin{cases} \min\limits_{\boldsymbol{U},\boldsymbol{V}} \mathrm{D}(\boldsymbol{X}||\boldsymbol{Z}) \\ \quad \boldsymbol{Z} = \boldsymbol{UV} \\ \boldsymbol{U} \geq 0 \quad \boldsymbol{V} \geq 0, \end{cases} \tag{11}$$

where $\mathrm{D}(\boldsymbol{X}||\boldsymbol{Z}) = \sum_{ij} \left(X_{ij} \log \frac{X_{ij}}{Z_{ij}} + X_{ij} - Z_{ij} \right)$.

In [27], the authors approximated the data matrix $\boldsymbol{X}$ that concatenates the pixel vectors from human face images. Each row of basis matrix $\boldsymbol{V}^{d\times D}$ can be considered as a basis image, which represents part of the human image, while each row of representation matrix $U^{N\times d}$ is the coefficient we can use to reconstruct the original human face images. Normally, $d << D$ indicates that representation matrix $U^{N\times d}$ is the desirable low-dimensional representation. To deal with outliers, Kong et al. [29] provided a robust NMF by enforcing the $\ell_{2,1}$ norm $\|\boldsymbol{X} - \boldsymbol{UV}\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{D} (\boldsymbol{X} - \boldsymbol{UV})_{j,i}^2}$, which is not squared, and thus, the large errors due to outliers do not dominate the objective function. There are many algorithms to solve this problem, like the classical multiplicative updates [30], projected gradient descent [31], coordinate descent [32], and the Alternating Direction Method of Multipliers (ADMM) [33].

Based on the above NMF, two groups of supervised NMF methods are proposed according to the means of using the label information. The first group introduced the loss function involving the label information into the objective function, while the second group borrowed the idea of LDA to improve the prediction ability of the obtained low-dimensional representation. We call them direct supervised NMF and discriminative NMF, respectively.

3.2.1. Direct Supervised NMF

In supervised learning like classification and regression, the label information is exploited in loss functions. Common loss functions for regression include quadratic loss, mean absolute error, and Huber loss, while common loss functions for classification include logistic loss, hinge loss, and KL divergence.

Lee et al. [34] integrated the quadratic loss into general NMF to form a semi-supervised NMF as:

$$\begin{cases} \min\limits_{\boldsymbol{U},\boldsymbol{V}} \|\boldsymbol{W} \odot (\boldsymbol{X} - \boldsymbol{UV})\|^2 + \alpha \|\boldsymbol{W} \odot (\boldsymbol{Y} - \boldsymbol{US})\|^2 \\ \qquad \boldsymbol{U} \geq 0 \quad \boldsymbol{V} \geq 0, \end{cases} \tag{12}$$

where α indicates the trade-off parameter. Assuming the number of classes is C, $\boldsymbol{Y} \in \mathbb{R}^{N \times C}$ denotes the label matrix. $\boldsymbol{W}$ is the indicator matrix that indicates whether $\boldsymbol{Y}_{ij}$ is observed, i.e.,

$$\boldsymbol{W}_{ij} = \begin{cases} 1 & \boldsymbol{Y}_{ij} \text{ is observed} \\ 0 & \boldsymbol{Y}_{ij} \text{ is missing.} \end{cases} \tag{13}$$

Based on [34], ref. [35] enforced an additional regularization to retain the difference of data points between different classes and formed their supervised NMF as:

$$\begin{cases} \min\limits_{\boldsymbol{U},\boldsymbol{V}} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}\|^2 + \alpha \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\|^2 + \beta \text{tr}(\boldsymbol{U}^T \boldsymbol{\Theta} \boldsymbol{U}) \\ \qquad \boldsymbol{U} \geq 0, \quad \boldsymbol{V} \geq 0, \quad \boldsymbol{\Theta} \geq 0, \end{cases} \tag{14}$$

where $\boldsymbol{\Theta}$ is an $N \times N$ matrix with each entry $\boldsymbol{\Theta}_{ij}$ equaling one if $y_i = y_j$ or zero otherwise, for $i, j = 1, \cdots, N$. β is the trade-off parameter. The introduction of the third item is to make the low-dimensional representations of data points in different classes differ greatly.

In order to combine NMF and the Support Vector Machine (SVM) classifier, Gupta and Xiao [36] proposed the general formulation for this problem as follows:

$$\begin{cases} \qquad \min\limits_{\boldsymbol{U},\boldsymbol{V},\boldsymbol{w}} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}\|^2 + \left(\|\boldsymbol{w}|\, \|^2 + C \sum_{i=1}^{N} L(y_i, \boldsymbol{w}^T \boldsymbol{u}_i + w_0) \right) \\ s.t. \qquad \boldsymbol{U} \in \mathcal{R}_+^{N \times d}, \boldsymbol{V} \in \mathbb{R}_+^{d \times D}, \boldsymbol{w} \in \mathbb{R}^{d \times 1}, y_i \in \{+1, -1\} \forall i, w_0 \in \mathbb{R} \end{cases} \tag{15}$$

where $(\boldsymbol{X}, \boldsymbol{Y})$ are the original data matrix and label vector. $\boldsymbol{Y}$ is composed of $y_i, i = 1, \cdots, N$. $\boldsymbol{V}$ is the basis matrix. $\boldsymbol{U}$ is the coefficient matrix. $\boldsymbol{u}_i$ is each row of $\boldsymbol{U}$. $L(\cdot, \cdot)$ is the loss function for the classifier. $\boldsymbol{w}$ and w_0 are the weight parameters and bias of the classifier, respectively. This type of supervised NMF can be considered as transforming the classification task from domain $(\boldsymbol{X}, \boldsymbol{Y})$ to $(\boldsymbol{U}, \boldsymbol{Y})$. Gupta and Xiao [36] adopted the loss function $L(y, t) = \max(0, 1 - yt)^p$, and p is a hyperparameter. It can be seen that when the margin yt is larger than one, there is no loss; this is a max-margin classifier. An alternative optimization strategy is then adopted to solve this problem.

Shu et al. [37] introduced multinomial loss into the framework (15) to deal with the multi-class classification problem. Chao et al. [38] integrated logistic loss and NMF into the unified framework explicitly and solved it with a projected gradient descent algorithm. They showed improved performance in predicting ICU 30-day mortality, compared with its unsupervised counterpart [39].

Mairal et al. [40,41] proposed a task-driven dictionary learning, which would become supervised NMF when requiring the dictionary and coefficient parameter to be nonnegative. Its main idea is integrating the dictionary learning and training of the classifier into a joint optimization problem, which is similar to that in [36]. Based on [41], Zhang et al. [42] enforced ℓ_1 regularization to make the new method robust to noises. To solve the acoustic separation problem, Bisot et al. [43] and Sprechmann et al. [44] made a modification by classifying the mean of the projections to adapt to the specific task.

3.2.2. Discriminative NMF

LDA aims to find a transformation to maximize the between-class distance and minimize the within-class distance. It is obviously a way to utilize the label information, and this idea was firstly reflected in [45] to conduct supervised NMF.

Let S_w and S_b measure the within-class and between-class scatter, respectively. Suppose there are C classes, and let n_i denote the number of vectors in the ith class.

$$S_w = \frac{1}{C} \sum_{i=1}^{C} \frac{1}{n_i} \sum_{j=1}^{n_i} (\boldsymbol{u}_j - \boldsymbol{m}_i)^T (\boldsymbol{u}_j - \boldsymbol{m}_i) \tag{16}$$

$$S_b = \frac{1}{C(C-1)} \sum_{i,j=1}^{C} \sum_{j=1}^{n_i} (\boldsymbol{m}_i - \boldsymbol{m}_j)^T (\boldsymbol{m}_i - \boldsymbol{m}_j) \tag{17}$$

where $\boldsymbol{m}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \boldsymbol{u}_j$ indicates the mean vector of class i in $\boldsymbol{U}$.

Based on the above concepts, Fisher (LDA is also called Fisher LDA) NMF [45] is formulated as:

$$\begin{cases} \min\limits_{\boldsymbol{U},\boldsymbol{V}} \mathrm{D}(\boldsymbol{X}||\boldsymbol{Z}) + \alpha S_w - \alpha S_b \\ \boldsymbol{Z} = \boldsymbol{U}\boldsymbol{V} \\ \boldsymbol{U} \geq 0 \quad \boldsymbol{V} \geq 0, \end{cases} \tag{18}$$

where α is the trade-off parameter. It can be noted that when $\alpha = 0$, it becomes the unsupervised NMF.

With the same idea, Zafeiriou et al. [46] and Kotsia et al. [47] provided another approach by setting different weights to the between-class and within-class scatter items instead of the same weight like α in Equation (18). Guan et al. [48] and Lu et al. [49] added more desirable properties such as smooth or discriminative in the basis matrix to discriminant NMF. Vilamala et al. [50] and Lee et al. [51] successfully applied discriminative NMF to human brain tumor classification and emotion classification.

3.3. Manifold-Based Supervised Dimension Reduction

Manifold learning assumes that the high-dimensional data points have a low-dimensional manifold, and the task of manifold learning is to uncover this low-dimensional manifold. Manifold-based dimension reduction methods exploit the geometric properties of the manifold on which the data points are supposed to lie. Common manifold-based dimension reduction methods include Isomap [52], Locally Linear Embedding (LLE) [53], and Laplacian Eigenmap (LE) [54]. We will introduce the above unsupervised manifold-based dimension reduction methods and their corresponding supervised versions in the following three subsections.

3.3.1. Isomap-Based Supervised Dimension Reduction

An earlier classical dimension reduction method, Multidimensional Scaling (MDS) [55], just retains the Euclidean distance and does not consider the neighborhood distribution, so it cannot deal with the case where high-dimensional data points lie on or near a curved manifold, like the Swiss roll dataset [52]. To overcome this drawback, Isomap attempts to preserve the pairwise geodesic distance that is measured on the manifold. It can be considered as the extension of MDS. To facilitate the understanding of supervised Isomap, we display algorithms of MDS and Isomap in Algorithms 1 and 2, respectively.

Algorithm 1 MDS algorithm.

```
Input: Distance matrix $\boldsymbol{D} \in \mathbb{R}^{N \times N}$, $d$
1. Calculate $\boldsymbol{B} = -\frac{1}{2}\boldsymbol{HDH}$, where $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^T$ is the centering matrix.
2. Conduct eigenvalue decomposition of $\boldsymbol{B}$ : $\boldsymbol{Bv} = \lambda \boldsymbol{v}$.
Output: $\boldsymbol{U} = \boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}$; $\boldsymbol{V}$ indicates the matrix of d eigenvectors, and $\boldsymbol{\lambda}$ is a diagonal matrix with diagonal entries as the largest deigenvalues.
```

Algorithm 2 Isomap algorithm.

```
Input: $x_1, \cdots, x_N \in \mathbb{R}^D$
1. Construct a graph with edge weight $\boldsymbol{W}_{ij} = ||x_i - x_j||$ for points $x_i$, $x_j$ in the $k$-nearest neighborhood or $\epsilon$-ball.
2. Compute the shortest distances between all pairs of points using Dijkstra's or Floyd's algorithm, and obtain the squares of the distances in matrix $\boldsymbol{D}$.
Output: MDS($\boldsymbol{D}$).
```

The work in [56] was the first to explore supervised Isomap by combining the Isomap procedure with the nearest neighbor classifier. Two supervised Isomap methods named WeightedIso and Iso+Ada, which took into consideration the label information by modifying the transformation performed by Isomap, were proposed in [56]. By designing the dissimilarity measures to integrate the label information, Ribeiro et al. [57] proposed an enhanced supervised Isomap. The dissimilarity measure [58] involved is defined as:

$$D(x_i, x_j) = \begin{cases} ((a-1)/a)^{1/2} & \text{if } c_i = c_j \\ a^{1/2} - d_0 & \text{if } c_i \neq c_j \end{cases} \tag{19}$$

where $a = 1/e^{-d_{ij}^2/\sigma}$ with d_{ij} set to be any distance measure, σ is a smoothing parameter, d_0 is a constant ($0 \leq d_0 \leq 1$), and c_i, c_j are the data class labels. The between-class dissimilarity is larger than the within-class dissimilarity, conferring a high discriminative power to this method.

Based on the above dissimilarity distance, the enhanced supervised Isomap is summarized in Algorithm 3.

Algorithm 3 Enhanced supervised Isomap.

```
Input: $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N \in \mathbb{R}^D, k, c_i, i = 1, 2$
1. Compute the dissimilarity matrix using label information from Equation (19).
2. Run Isomap in Algorithm 2 to obtain low embedding map $\boldsymbol{U}$.
3. Learn the embedded mapping $\boldsymbol{D}$ to construct dissimilarity kernels.
4. SVM tests on new points.
Output: $\boldsymbol{D}$
```

Li and Guo [59] not only obtained explicit mapping from high-dimensional space to low-dimensional space during supervised Isomap learning, but also adopted geodesic distance instead of Euclidean distance to make this Isomap robust to noise. To exploit the labeled and unlabeled data points, Zhang et al. [60] provided a semi-supervised Isomap by mining the pairwise within-class distances in the same manifold and maximizing the distances between different manifolds.

3.3.2. LLE-Based Supervised Dimension Reduction

In contrast with Isomap, which retains the global structure property, LLE attempts to preserve the local structure property. It assumes that each data point in the original space can be represented as a linear combination of their nearest neighbors, and it tries to look for the low-dimensional representations of these data points to keep this linear combination property.

Suppose that a data point $\boldsymbol{x}_i$ can be written as a linear combination w_{ij} of its k nearest neighbors $\boldsymbol{x}_j$. Note that the k nearest neighbors are identified by ranking the dissimilarity matrix $\boldsymbol{\Delta}$. The LLE can be formulated as the following optimization problem.

$$\begin{cases} \min\limits_{\boldsymbol{U}} \sum_{i=1}^{N} \left\| \boldsymbol{u}_i - \sum_{j=1}^{k} w_{ij} \boldsymbol{u}_j \right\|^2 \\ s.t. \frac{1}{N} \boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{I}. \end{cases} \tag{20}$$

where u^k indicates the k column of the solution matrix $\boldsymbol{U}$. The constraint is enforced to avoid the trivial solution $\boldsymbol{U} = \boldsymbol{0}$.

By modifying the dissimilarity, De Ridder and Duin [61] and De Ridder et al. [62] proposed the supervised LLE. The modified dissimilarity matrix $\boldsymbol{\Delta}^{'} = \boldsymbol{\Delta} + \alpha \max(\boldsymbol{\Delta})$ where $0 \leq \alpha \leq 1$, $\max(\boldsymbol{\Delta})$ is the maximum entry of $\boldsymbol{\Delta}$ and $\Lambda_{ij} = 1$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ belong to the same class, and zero otherwise. Obviously, when $\alpha = 0$, it becomes unsupervised LLE. When $\alpha = 1$, it is the fully-supervised LLE, and when $0 < \alpha < 1$, it is the semi-supervised LLE. After modifying the dissimilarity matrix, all the subsequent steps are the same as for LLE.

Zhang [63] and Liu et al. [64] adopted the same idea that the between-class dissimilarity is larger than within-class dissimilarity to conduct supervised LLE. Moreover, Liu et al. [64] extended supervised LLE in tensor space to handle high order data to retain their structure information in each order. We can sum up that all these supervised LLE methods reflect the LDA idea.

3.3.3. LE-Based Supervised Dimension Reduction

LE [54] attempts to preserve the local neighborhood structure by using the Laplacian of the graph. The similarity matrix can be constructed by using Gaussian function $W_{ij} = \exp(-\frac{||x_i - x_j||^2}{\beta})$ where $i, j = 1, \cdots, N, \beta$ is a scale parameter that is usually set to the average of squared distances between all pairs. LE tries to seek the low-dimensional representation $u_i, i = 1, \cdots, N$ by minimizing $-\frac{1}{2} \sum_{i,j} ||x_i - x_j||^2 W_{i,j} = \mathrm{tr}(\boldsymbol{U}^T \boldsymbol{L} \boldsymbol{U})$. Therefore, LE can be formulated as:

$$\begin{cases} \min\limits_{\boldsymbol{U}} \mathrm{tr}(\boldsymbol{U}^T \boldsymbol{L} \boldsymbol{U}) \\ s.t. \ \boldsymbol{U}^T \boldsymbol{D} \boldsymbol{U} = \boldsymbol{I}, \\ \quad \boldsymbol{U}^T \boldsymbol{L} \boldsymbol{e} = \boldsymbol{0}. \end{cases} \tag{21}$$

where $\boldsymbol{I}$ is the identity matrix and $\boldsymbol{e} = (1, \cdots, 1)^T$, $\boldsymbol{D}$ is the diagonal matrix whose entries are column or row sums of similarity matrix $\boldsymbol{W}$, $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$ is the Laplacian matrix, and $\boldsymbol{U}$ is the low-dimensional matrix we seek. The two constraints in Equation (21) are used to avoid the trivial solutions $\boldsymbol{U} = \boldsymbol{0}$ and $\boldsymbol{U} = \boldsymbol{e}$. Applying the Lagrange multiplier method and using the fact $\boldsymbol{L}\boldsymbol{e} = \boldsymbol{0}$, the solutions of Equation (21) can be obtained by forming a matrix by the eigenvectors corresponding to the smallest deigenvalues (excluding zero) of the generalized eigenvector problem as:

$$\boldsymbol{L}\boldsymbol{u} = \lambda \boldsymbol{D}\boldsymbol{u}. \tag{22}$$

In order to adapt LE for the classification task, borrowing the idea of LDA, Raducanu and Dornaika [65] proposed a supervised LE. By minimizing the margin between homogeneous data points and maximizing the margin between heterogeneous data points, supervised LE [65] exploited the label information well and learned the supervised low-dimensional representation finally. To define the margin, for each data point $\boldsymbol{x}_i$, they defined two sets $N_w(\boldsymbol{x}_i)$ and $N_b(\boldsymbol{x}_i)$ to indicate the within-class neighbors and between-class neighbors with a similarity higher than the average one, respectively.

$$N_w(\boldsymbol{x}_i) = \{\boldsymbol{x}_j | y_j = y_i, \exp(-\frac{||x_i - x_j||^2}{\beta}) > AS(\boldsymbol{x}_i)\} \tag{23}$$

$$N_w(\boldsymbol{x}_i) = \{\boldsymbol{x}_j | y_j \neq_i, \exp(-\frac{||x_i - x_j||^2}{\beta}) > AS(\boldsymbol{x}_i)\} \tag{24}$$

where $AS(\boldsymbol{x}_i) = \frac{1}{N}\sum_{k=1}^{N}\exp(-\frac{||x_i - x_j||^2}{\beta})$ indicates the average similarity of the sample $\boldsymbol{x}_i$ to all the rest of the data points.

With these two sets defined, two weight matrices corresponding to Equations (23) and (24) are defined as:

$$W_{w,ij} = \begin{cases} \exp(-\frac{||x_i - x_j||^2}{\beta}) & \text{if } \boldsymbol{x}_j \in N_w(\boldsymbol{x}_i) \text{ or } \boldsymbol{x}_i \in N_w(\boldsymbol{x}_j) \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

$$W_{w,ij} = \begin{cases} 1 & \text{if } \boldsymbol{x}_j \in N_b(\boldsymbol{x}_i) \text{ or } \boldsymbol{x}_i \in N_b(\boldsymbol{x}_j) \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

To get the low-dimensional representation $\boldsymbol{U}$, two objective functions can be optimized as follows:

$$\min \frac{1}{2}\sum_{i,j} ||\boldsymbol{u}_i - \boldsymbol{u}_j||^2 W_{w,ij} = \mathrm{tr}(\boldsymbol{U}^T\boldsymbol{L}_w\boldsymbol{U}) \tag{27}$$

$$\max \frac{1}{2}\sum_{i,j} ||\boldsymbol{u}_i - \boldsymbol{u}_j||^2 W_{b,ij} = \mathrm{tr}(\boldsymbol{U}^T\boldsymbol{L}_b\boldsymbol{U}) \tag{28}$$

where $\boldsymbol{L}_w = \boldsymbol{D}_w - \boldsymbol{W}_w$ and $\boldsymbol{L}_b = \boldsymbol{D}_b - \boldsymbol{W}_b$ indicate the corresponding Laplacians.

By merging the above two objective functions, the final optimization problem is formulated as:

$$\begin{cases} \max\limits_{\boldsymbol{U}} \mathrm{tr}(\boldsymbol{U}^T\boldsymbol{L}_b\boldsymbol{U}) + (1-\gamma)\mathrm{tr}(\boldsymbol{U}^T\boldsymbol{L}_w\boldsymbol{U}) \\ \qquad s.t.\ \boldsymbol{U}^T\boldsymbol{D}_w\boldsymbol{U} = \boldsymbol{I}. \end{cases} \tag{29}$$

By defining matrix $\boldsymbol{B} = \gamma\boldsymbol{L}_b + (1-\gamma)\boldsymbol{W}_w$, the above problem can be transformed as:

$$\begin{cases} \max\limits_{\boldsymbol{U}} \mathrm{tr}(\boldsymbol{U}^T\boldsymbol{B}\boldsymbol{U}) \\ s.t.\ \boldsymbol{U}^T\boldsymbol{D}_w\boldsymbol{U} = \boldsymbol{I}. \end{cases} \tag{30}$$

This formulation is easy to solve by the generalized eigenvalue problem.

Besides the above popular supervised LE method, Zheng et al. [66] explored another way to integrate the label information by optimizing the weight matrix using the labels after constructing the similarity matrix from local neighborhood relation. Wu et al. [67] proposed a deep learning-based supervised LE method whose deep architecture consists of multiple stacked layers and computes an intermediate representation that is fed to a nearest-neighbor classifier. Jiang and Jia [68] integrated the label information into the process of constructing the dissimilarity matrix, and the other steps are the same as for the general LE.

3.4. Discussion

In the three introduced classes of supervised or semi-supervised dimension reduction methods, supervised NMF has been successfully applied in computer vision and speech recognition, because NMF has a very good interpretability due to its non-negativity property. PCA-based methods can be used in all the classification or regression problems, but their performance may not be as competitive as NMF-based methods in the computer vision and speech recognition fields. Manifold-based methods assume that the data points are located in a low-dimensional manifold or each data point can be represented as the linear combination of its neighbors; thus, they are not as general as PCA-based method, but more general than NMF-based methods. In addition, manifold-based methods are normally time consuming due to the inverse of the Laplacian matrix. In summary, from the perspective of generality, the three classes of supervised or semi-supervised methods are ranked as PCA-based methods, manifold-based methods, and then, NMF-based methods.

4. Application

Supervised dimension reduction has been successfully applied to a variety of applications including computer vision, biomedical informatics, speech recognition, visualization, etc.

4.1. Computer Vision

From the inception of NMF [27], it had been successfully applied to face recognition due to its ability to produce interpretable bases. Naturally, Face recognition becomes the typically successful application of supervised NMF. Discriminative NMFs [46,47,69] are the earlier successful attempts of supervised NMF methods at face recognition, and then, many direct NMF methods [35–37,70] also demonstrated superior performance in this task.

Apart from face recognition, all this object or action recognition also involves the application of supervised dimension reduction. Wu et al. [67] proposed a supervised Laplacian eigenmap to recognize visual objects. Kumar [71] adopted supervised dictionary learning to recognize the actions and locations of the objects in the images. Santiago-Mozos et al. [72] applied supervised PCA to object detection in infrared images and demonstrated good performance. Recently, Xinfang et al. [73] proposed a semi-supervised local discriminant analysis by combing the idea of LDA and LLE for polarimetric SAR image classification.

4.2. Biomedical Informatics

In bioinformatics, especially genetics, due to the large amount of gene markers, it is challenging to identify the true gene marker that results in a certain disease directly. Two tough goals, high dimension and classification, should be simultaneously tackled; thus, supervised dimension reduction becomes the ideal choice. Zhang et al. [74] proposed a semi-supervised projective NMF method for cancer classification. Gaujoux and Seoighe [75] adopted another semi-supervised NMF method for gene expression deconvolution. Supervised PCA [76] was successfully applied to gene set analysis, while supervised categorical PCA [77,78] was successfully applied in genome-wide association analyses. Moreover, supervised probabilistic PCA [26] performed rather well in gene classification.

In medical informatics, with the fast development of medical devices, a variety of features are collected in real applications. Inevitably, some noisy, redundant, or useless features are included, which hinders identifying certain diseases. How to identify the effective features for certain diseases is challenging, and supervised dimension reduction becomes a good option to solve this problem. Vilamala et al. [50] designed a discriminative NMF and successfully applied ti to human brain tumor classification. Chao et al. [38] proposed a supervised NMF by combing NMF and logistic regression and improved the ICU mortality prediction performance. Fuse et al. [79] combined NMF and SVM to diagnose Alzheimer's disease and obtained an improved performance. Supervised PCA [20] has been successfully used in DNA microarray data analysis and cancer diagnosis. It is noted that the process of knowledge discovery in biomedical informatics is mostly performed by biomedical domain experts. This is mostly due to the high complexity of the research domain, which requires deep domain knowledge. At the same time, these domain experts face major obstacles in handling and analyzing their high-dimensional, heterogeneous, and complex research data. A recent work [80] outlined that ontology-centered data infrastructure for scientific research, which actively supports the medical domain experts in data acquisition, processing, and exploration, can be very beneficial here.

4.3. Speech Recognition

Speech recognition is another successful application of NMF, and thus, supervised NMF is naturally successfully used in this kind of application. Lee et al. [51] used discriminative NMF to classify the emotional difference in speech. Bisot et al. [43] applied supervised NMF to acoustic scene classification and obtained rather good performance. Sprechmann et al. [44] and

Weninger et al. [81] solved the audio source separation with supervised NMF, while Nakajima et al. [82] and Kitamura et al. [83] adopted supervised NMF for music signal separation.

Although there exist an amount of successful applications in speech recognition, more attempts can be made in the future. As we can see that almost all of the existing supervised dimension reduction methods are NMF-based, both PCA-based and manifold-based methods can be investigated and compared with the existing methods.

4.4. Visualization

High-dimensional data are hard to explain. Take the ICU mortality prediction problem [38] as an example: there are many vital sign features, and it is difficult to interpret them individually due to the high dimensionality. As far as we know, biomedical experts are increasingly confronted with complex high-dimensional data. As the number of dimensions is often very large, one needs to map them to a smaller number of relevant dimensions to be more amenable to expert analysis. This is because irrelevant, redundant, and conflicting dimensions can negatively affect the effectiveness and efficiency of the analytic process. This is also the so-called curse of dimensionality problem. To deal with this problem, dimension reduction is a possible means, but the possible mappings from high- to low-dimensional spaces are ambiguous. Subspace analysis [84,85] can be used to seek solutions. Since high-dimensional data are difficult to interpret, a rough picture of the data is quite helpful; thus, visualization is very important, and it is also an important application of supervised dimension reduction. Barshan et al. [21] provided a supervised PCA to conduct visualization, while Vlachos et al. [56] gave another supervised dimension reduction method by borrowing the LDA idea for visualization. Geng et al. [58] proposed a supervised Isomap to visualize. Compared with visualization from general unsupervised dimension reduction, visualization from supervised dimension reduction has clear separability due to its supervised learning property.

Apart from all the above applications, text mining is probably another good application of supervised dimension reduction. Although there are already many works [86–88] on unsupervised dimension reduction, there are few works on supervised dimension reduction.

5. Potential Future Research Issues

Although supervised dimension reduction has developed greatly and been successfully applied to many applications during the last two decades, there are still some challenging problems that need to be tackled in the future. Below, we unfold some important open problems worth further exploration.

5.1. Scalability

For PCA-based methods, the time complexity of covariance matrix computation is $O(D^2N)$, and that of its eigenvalue decomposition is $O(D^3)$. Therefore, the complexity of PCA is $O(D^2N + D^3)$. For NMF-based methods, some fast solving methods like the projected gradient descent method [31] do not work due to the additional objective function items, then the time complexity of its most time-costly part is $O(tNDd)$; t is the iteration numbers it needs to converge. For manifold-based methods, the time complexity of constructing the similarity matrix is $O(N^2D)$, and the frequently-used solving strategy is generalized eigenvalue decomposition; the time complexity is $O(D^3)$. One of the main goals of supervised dimension reduction is to solve high-dimensional problems, but when the feature dimension is high, the time costs of the existing supervised dimension reduction methods are still high, because some specifically-designed unsupervised dimension reduction methods do not work due to the appearance of new objective items or constraints on label information. When dataset is huge in sample size, like in social networks, there are millions of data points, and the time cost for supervised dimension reduction is still unacceptable. Therefore, some specific algorithms directed at supervised dimension reduction are urgently in need, especially due to the data explosion in this era.

5.2. *Missing Values*

Missing values are a common phenomenon in many applications due to a variety of factors like the failure of sensors in computer vision and missing certain laboratory test results over time for some patients in the clinical setting [89]. The existing strategy is imputation with zero, the mean, or the maximum value, or multiple imputation [90]. In order to tackle missing values, Lee et al. [34] introduced an auxiliary matrix to indicate whether the entry was missed or not. Obviously, no specific designs are involved in the supervised dimension reduction process. Some tricks to handle missing values like the E-M algorithm [91] can be considered to be incorporated into some supervised dimension reduction methods. In addition, multi-view information of the data has consensus, and they are complementary to each other [92–94], which can be the other direction to handle the missing value problem.

5.3. *Heterogeneous Types*

Data may contain heterogeneous types of features such as numerical, categorical, symbolic, ordinal features, etc. How to integrate different types of data together to perform supervised dimension reduction is a challenging problem. A natural way to handle this problem is to convert all of them to the categorical type. However, much information will be lost during this phase. For instance, the difference of the continuous values categorized into the same category is ignored [95]. Therefore, how to exploit the information within mixed data types is worth exploring in the future.

Besides the above three potential research issues, an emerging future research issue that will become very important in the future is the explanation part, and this will require supervised dimension reduction to make results from arbitrarily high-dimensional spaces understandable for a human, who can perceive information only in the lower dimensions. We can refer to the recent work [96] to learn about this direction. Apart from supervised dimension reduction, it is also intriguing to explore other ways to explain high-dimensional data well.

6. Conclusions

The field of supervised dimension reduction has seen extensive growth at an increasing rate. We have outlined the state-of-the-art research in this review by categorizing it into three main classes: PCA-based, NMF-based, and manifold-based supervised dimension reduction methods. To understand their characteristics better, we provide a discussion to elaborate their advantages and disadvantages. To advance the further development of this topic, we also list some open problems waiting for analytical study in the future. This review will be helpful for researchers who want to develop advanced supervised dimension reduction methods or who seek methods to learn low-dimensional representation for certain supervised learning applications. We believe that supervised dimension reduction will continue to remain an active area of study in the years to come, owing to an increase in the high-dimensional data and sustained community efforts. In addition, their tighter integration into specific application systems will continuously shape the emerging landscape and provide opportunities for researcher contribution.

Author Contributions: G.C. created the structure and organization of this paper and wrote the first draft of this paper. W.D. improved the writing and advised to add a Section 3 to introduce applications and unify the formulation. Y.L. improved the writing and advised to create Section 3.4 to introduce visualization in a separate subsection and improve Figure 1. Both Y.L. and W.D. contributed to the final version of the manuscript. All authors approved of the final version of the manuscript.

Funding: Weiping Ding is funded by the National Natural Science Foundation of China under Grant 61300167 and the Six Talent Peaks Project of Jiangsu Province under Grant XYDXXJS-048.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; nor in the decision to publish the results.

References

1. Ai-Jun, Y.; Xin-Yuan, S. Bayesian variable selection for disease classification using gene expression data. *Bioinformatics* **2009**, *26*, 215–222. [CrossRef]
2. Sun, J.; Bi, J.; Kranzler, H.R. Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC Genet.* **2014**, *15*, 73. [CrossRef]
3. Luo, Y.; Mao, C.; Yang, Y.; Wang, F.; Ahmad, F.S.; Arnett, D.; Irvin, M.R.; Shah, S.J. Integrating Hypertension Phenotype and Genotype with Hybrid Non-negative Matrix Factorization. *Bioinformatics* **2018**.[CrossRef]
4. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. [CrossRef]
5. Sun, S.; Zhang, C. Adaptive feature extraction for EEG signal classification. *Med. Biol. Eng. Comput.* **2006**, *44*, 931–935. [CrossRef]
6. Guyon, I.; Elisseeff, A. An introduction to feature extraction. In *Feature Extraction*; Springer: Berlin/Heidelberg, German, 2006; pp. 1–25.
7. Rogati, M.; Yang, Y. High-performing feature selection for text classification. In Proceedings of the Eleventh International Conference on Information and Knowledge Management, McLean, VA, USA, 4–9 November 2002; pp. 659–661.
8. Kim, H.; Howland, P.; Park, H. Dimension reduction in text classification with support vector machines. *J. Mach. Learn. Res.* **2005**, *6*, 37–53.
9. Basu, T.; Murthy, C. Effective text classification by a supervised feature selection approach. In Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops, Brussels, Belgium, 10 December 2012; pp. 918–925.
10. Carreira-Perpinán, M.A. *A Review of Dimension Reduction Techniques*; Technical Report CS-96-09 9; University of Sheffield: Sheffield, UK, 1997, pp. 1–69.
11. Fodor, I.K. *A Survey of Dimension Reduction Techniques*; Center for Applied Scientific Computing, Lawrence Livermore National Laboratory: Livermore, CA, USA, 2002, Volume 9,pp. 1–18.
12. Van Der Maaten, L.; Postma, E.; Van den Herik, J. Dimensionality reduction: A comparative review. *J. Mach. Learn. Res.* **2009**, *10*, 66–71.
13. Thangavel, K.; Pethalakshmi, A. Dimensionality reduction based on rough set theory: A review. *Appl. Soft Comput.* **2009**, *9*, 1–12. [CrossRef]
14. Ma, Y.; Zhu, L. A review on dimension reduction. *Int. Stat. Rev.* **2013**, *81*, 134–150. [CrossRef]
15. Blum, M.G.; Nunes, M.A.; Prangle, D.; Sisson, S.A. A comparative review of dimension reduction methods in approximate Bayesian computation. *Stat. Sci.* **2013**, *28*, 189–208. [CrossRef]
16. Sorzano, C.O.S.; Vargas, J.; Montano, A.P. A survey of dimensionality reduction techniques. *arXiv* **2014**, arXiv:1403.2877.
17. Luo, Y.; Ahmad, F.S.; Shah, S.J. Tensor factorization for precision medicine in heart failure with preserved ejection fraction. *J. Cardiovasc. Transl. Res.* **2017**, *10*, 305–312. [CrossRef] [PubMed]
18. Tang, J.; Alelyani, S.; Liu, H. A survey of dimensionality reduction techniques. In *Data Classification: Algorithms and Applications*; CRC Press: Boca Raton, FL, USA, 2015.
19. Hotelling, H. Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **1933**, *24*, 417. [CrossRef]
20. Bair, E.; Hastie, T.; Paul, D.; Tibshirani, R. Prediction by supervised principal components. *J. Am. Stat. Assoc.* **2006**, *101*, 119–137. [CrossRef]
21. Barshan, E.; Ghodsi, A.; Azimifar, Z.; Jahromi, M.Z. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognit.* **2011**, *44*, 1357–1371. [CrossRef]
22. Gretton, A.; Bousquet, O.; Smola, A.; Schölkopf, B. Measuring statistical dependence with Hilbert-Schmidt norms. In *International Conference on Algorithmic Learning Theory*; Springer: Berlin/Heidelberg, German, 2005; pp. 63–77.
23. Fukumizu, K.; Bach, F.R.; Jordan, M.I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **2004**, *5*, 73–99.
24. Bin, J.; Ai, F.F.; Liu, N.; Zhang, Z.M.; Liang, Y.Z.; Shu, R.X.; Yang, K. Supervised principal components: A new method for multivariate spectral analysis. *J. Chemom.* **2013**, *27*, 457–465. [CrossRef]

25. Roberts, S.; Martin, M.A. Using supervised principal components analysis to assess multiple pollutant effects. *Environ. Health Perspect.* **2006**, *114*, 1877. [CrossRef]
26. Yu, S.; Yu, K.; Tresp, V.; Kriegel, H.P.; Wu, M. Supervised probabilistic principal component analysis. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Philadelphia, PA, USA, 20–23 August 2006; pp. 464–473.
27. Lee, D.D.; Seung, H.S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788. [CrossRef]
28. Dhillon, I.S.; Tropp, J.A. Matrix nearness problems with Bregman divergences. *SIAM J. Matrix Anal. Appl.* **2007**, *29*, 1120–1146. [CrossRef]
29. Kong, D.; Ding, C.; Huang, H. Robust nonnegative matrix factorization using l21-norm. In Proceedings of the 20th ACM International Conference on Information and Knowledge Management, Glasgow, UK, 24–28 October 2011; pp. 673–682.
30. Lee, D.D.; Seung, H.S. Algorithms for non-negative matrix factorization. In proceedings of the Conference on Neural Information Processing Systems, Denver, CO, USA, 27 November–2 December 2000; pp. 556–562.
31. Lin, C.J. Projected gradient methods for nonnegative matrix factorization. *Neural Comput.* **2007**, *19*, 2756–2779. [CrossRef] [PubMed]
32. Hsieh, C.J.; Dhillon, I.S. Fast coordinate descent methods with variable selection for non-negative matrix factorization. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, 21–24 August 2011; pp. 1064–1072.
33. Sun, D.L.; Fevotte, C. Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence. In Proceedings of the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 6201–6205.
34. Lee, H.; Yoo, J.; Choi, S. Semi-supervised nonnegative matrix factorization. *IEEE Signal Process. Lett.* **2010**, *17*, 4–7.
35. Jing, L.; Zhang, C.; Ng, M.K. SNMFCA: Supervised NMF-based image classification and annotation. *IEEE Trans. Image Process.* **2012**, *21*, 4508–4521. [CrossRef] [PubMed]
36. Gupta, M.D.; Xiao, J. Non-negative matrix factorization as a feature selection tool for maximum margin classifiers. In Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Colorado Springs, CO, USA, 20–25 June 2011; pp. 2841–2848.
37. Shu, X.; Lu, H.; Tao, L. Joint learning with nonnegative matrix factorization and multinomial logistic regression. In Proceedings of the 2013 International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; pp. 3760–3764.
38. Chao, G.; Mao, C.; Wang, F.; Zhao, Y.; Luo, Y. Supervised Nonnegative Matrix Factorization to Predict ICU Mortality Risk. *arXiv* **2018**, arXiv:1809.10680 .
39. Luo, Y.; Xin, Y.; Joshi, R.; Celi, L.A.; Szolovits, P. Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 42–50.
40. Mairal, J.; Bach, F.; Ponce, J.; Sapiro, G.; Zisserman, A. Discriminative learned dictionaries for local image analysis. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008.
41. Mairal, J.; Bach, F.; Ponce, J. Task-driven dictionary learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 791–804. [CrossRef] [PubMed]
42. Zhang, R.; Hu, Z.; Pan, G.; Wang, Y. Robust discriminative non-negative matrix factorization. *Neurocomputing* **2016**, *173*, 552–561. [CrossRef]
43. Bisot, V.; Serizel, R.; Essid, S.; Richard, G. Supervised nonnegative matrix factorization for acoustic scene classification. In Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016, Budapest, Hungary, 3 September 2016.
44. Sprechmann, P.; Bronstein, A.M.; Sapiro, G. Supervised non-negative matrix factorization for audio source separation. In *Excursions in Harmonic Analysis*; Springer: Berlin/Heidelberg, Germany, 2015; Volume 4, pp. 407–420.
45. Wang, Y.; Jia, Y.; Hu, C.; Turk, M. Fisher non-negative matrix factorization for learning local features. In Proceedings of the Sixth Asian Conference on Computer Vision, Jeju, Korea, 27–30 January 2004; pp. 27–30.

46. Zafeiriou, S.; Tefas, A.; Buciu, I.; Pitas, I. Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification. *IEEE Trans. Neural Netw.* **2006**, *17*, 683–695. [CrossRef]
47. Kotsia, I.; Zafeiriou, S.; Pitas, I. A novel discriminant non-negative matrix factorization algorithm with applications to facial image characterization problems. *IEEE Trans. Inf. Forensics Secur.* **2007**, *2*, 588–595. [CrossRef]
48. Guan, N.; Tao, D.; Luo, Z.; Yuan, B. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. *IEEE Trans. Image Process.* **2011**, *20*, 2030–2048. [CrossRef]
49. Lu, Y.; Lai, Z.; Xu, Y.; Li, X.; Zhang, D.; Yuan, C. Nonnegative discriminant matrix factorization. *IEEE Trans. Circuits Syst. Video Technol.* **2017**, *27*, 1392–1405. [CrossRef]
50. Vilamala, A.; Lisboa, P.J.; Ortega-Martorell, S.; Vellido, A. Discriminant Convex Non-negative Matrix Factorization for the classification of human brain tumours. *Pattern Recognit. Lett.* **2013**, *34*, 1734–1747. [CrossRef]
51. Lee, S.Y.; Song, H.A.; Amari, S.I. A new discriminant NMF algorithm and its application to the extraction of subtle emotional differences in speech. *Cognit. Neurodyn.* **2012**, *6*, 525–535. [CrossRef] [PubMed]
52. Tenenbaum, J.B.; De Silva, V.; Langford, J.C. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323. [CrossRef] [PubMed]
53. Roweis, S.T.; Saul, L.K. Nonlinear dimensionality reduction by locally linear embedding. *Science* **2000**, *290*, 2323–2326. [CrossRef] [PubMed]
54. Belkin, M.; Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **2003**, *15*, 1373–1396. [CrossRef]
55. Torgerson, W.S. Multidimensional scaling: I. Theory and method. *Psychometrika* **1952**, *17*, 401–419. [CrossRef]
56. Vlachos, M.; Domeniconi, C.; Gunopulos, D.; Kollios, G.; Koudas, N. Non-linear dimensionality reduction techniques for classification and visualization. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 645–651.
57. Ribeiro, B.; Vieira, A.; das Neves, J.C. Supervised Isomap with dissimilarity measures in embedding learning. In *Iberoamerican Congress on Pattern Recognition*; Springer: Berlin/Heidelberg, German, 2008; pp. 389–396.
58. Geng, X.; Zhan, D.C.; Zhou, Z.H. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Trans. Syst. Man Cybern. Part B* **2005**, *35*, 1098–1107. [CrossRef]
59. Li, C.G.; Guo, J. Supervised isomap with explicit mapping. In Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06), Beijing, China, 30 August–1 September 2006; pp. 345–348.
60. Zhang, Y.; Zhang, Z.; Qin, J.; Zhang, L.; Li, B.; Li, F. Semi-supervised local multi-manifold Isomap by linear embedding for feature extraction. *Pattern Recognit.* **2018**, *76*, 662–678. [CrossRef]
61. De Ridder, D.; Duin, R.P. *Locally Linear Embedding for Classification*; Pattern Recognition Group Technical Report PH-2002-01; Delft University of Technology: Delft, The Netherlands, 2002; pp. 1–12.
62. De Ridder, D.; Kouropteva, O.; Okun, O.; Pietikäinen, M.; Duin, R.P. Supervised locally linear embedding. In *Artificial Neural Networks and Neural Information Processing—ICANN/ICONIP 2003*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 333–341.
63. Zhang, S.Q. Enhanced supervised locally linear embedding. *Pattern Recognit. Lett.* **2009**, *30*, 1208–1218. [CrossRef]
64. Liu, C.; Zhou, J.; He, K.; Zhu, Y.; Wang, D.; Xia, J. Supervised locally linear embedding in tensor space. In Proceedings of the 2009 Third International Symposium on Intelligent Information Technology Application, NanChang, China, 21–22 November 2009; pp. 31–34.
65. Raducanu, B.; Dornaika, F. A supervised non-linear dimensionality reduction approach for manifold learning. *Pattern Recognit.* **2012**, *45*, 2432–2444. [CrossRef]
66. Zheng, F.; Chen, N.; Li, L. Semi-supervised Laplacian eigenmaps for dimensionality reduction. In Proceedings of the 2008 International Conference on Wavelet Analysis and Pattern Recognition, Hong Kong, China, 30–31 August 2008; pp. 843–849.
67. Wu, R.; Yu, Y.; Wang, W. Scale: Supervised and cascaded laplacian eigenmaps for visual object recognition based on nearest neighbors. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 25–27 June 2013; pp. 867–874.

68. Jiang, Q.; Jia, M. Supervised laplacian eigenmaps for machinery fault classification. In Proceedings of the 2009 WRI World Congress on Computer Science and Information Engineering, Los Angeles, CA, USA, 31 March–2 April 2009; pp. 116–120.
69. Zhang, X.; Yali, P.; Liu, S.; Wu, J.; Ren, P. A supervised dimensionality reduction method-based sparse representation for face recognition. *J. Mod. Opt.* **2017**, *64*, 799–806. [CrossRef]
70. Chen, W.S.; Zhao, Y.; Pan, B.; Chen, B. Supervised kernel nonnegative matrix factorization for face recognition. *Neurocomputing* **2016**, *205*, 165–181. [CrossRef]
71. Kumar, B. Supervised Dictionary Learning for Action Recognition and Localization. Ph.D. Thesis, Queen Mary University of London, London, UK, 2012.
72. Santiago-Mozos, R.; Leiva-Murillo, J.M.; Pérez-Cruz, F.; Artes-Rodriguez, A. Supervised-PCA and SVM classifiers for object detection in infrared images. In Proceedings of the IEEE Conference on Advanced Video and Signal Based Surveillance, Miami, FL, USA, 22–22 July 2003; pp. 122–127.
73. Xinfang, X.; Xin, X.; Hao, D.; Han, W.; Luoru, L. A Semi-Supervised Dimension Reduction Method for Polarimetric SAR Image Classification. *Acta Opt. Sin.* **2018**, *4*, 045.
74. Zhang, X.; Guan, N.; Jia, Z.; Qiu, X.; Luo, Z. Semi-supervised projective non-negative matrix factorization for cancer classification. *PLoS ONE* **2015**, *10*, 1–20. [CrossRef]
75. Gaujoux, R.; Seoighe, C. Semi-supervised Nonnegative Matrix Factorization for gene expression deconvolution: a case study. *Infect. Genet. Evol.* **2012**, *12*, 913–921. [CrossRef]
76. Chen, X.; Wang, L.; Smith, J.D.; Zhang, B. Supervised principal component analysis for gene set enrichment of microarray data with continuous or survival outcomes. *Bioinformatics* **2008**, *24*, 2474–2481. [CrossRef]
77. Lu, M.; Lee, H.S.; Hadley, D.; Huang, J.Z.; Qian, X. Supervised categorical principal component analysis for genome-wide association analyses. *BMC Genom.* **2014**, *15*, 1–10. [CrossRef]
78. Lu, M.; Huang, J.Z.; Qian, X. Supervised logistic principal component analysis for pathway based genome-wide association studies. In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, Orlando, FL, USA, 7–10 October 2012; pp. 52–59.
79. Fuse, T.A.; Jayasignpure, N.D.; Pawar, P.D. NMF-SVM Based CAD Tool for the Diagnosis of Alzheimer's Disease. *Int. J. Appl. Innov. Eng. Manag.* **2014**, *3*, 268–274.
80. Giradi, D.; Holzinger, A. Dimensionality Reduction for Exploratory Data Analysis in Daily Medical Research. In *Advanced Data Analytics in Health*; Springer: Cham, Switzerland, 2018; pp. 3–20.
81. Weninger, F.; Roux, J.L.; Hershey, J.R.; Watanabe, S. Discriminative NMF and its application to single-channel source separation. In Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
82. Nakajima, H.; Kitamura, D.; Takamune, N.; Koyama, S.; Saruwatari, H.; Ono, N.; Takahashi, Y.; Kondo, K. Music signal separation using supervised NMF with all-pole-model-based discriminative basis deformation. In Proceedings of the 2016 24th European Signal Processing Conference (EUSIPCO), Budapest, Hungary, 29 August–2 September 2016; pp. 1143–1147.
83. Kitamura, D.; Saruwatari, H.; Yagi, K.; Shikano, K.; Takahashi, Y.; Kondo, K. Robust music signal separation based on supervised nonnegative matrix factorization with prevention of basis sharing. In Proceedings of the IEEE International Symposium on Signal Processing and Information Technology, Athens, Greece, 12–15 December 2013; pp. 392–397.
84. Hund, M.; Böhm, D.; Sturm, W.; Sedlmair, M.; Schreck, T.; Ullrich, T.; Keim, D.A.; Majnaric, L.; Holzinger, A. Visual analytics for concept exploration in subspaces of patient groups. *Brain Inform.* **2016**, *3*, 233–247. [CrossRef]
85. Sun, S.; Zhang, C. The selective random subspace predictor for traffic flow forecasting. *IEEE Trans. Int. Transp. Syst.* **2007**, *8*, 367–373. [CrossRef]
86. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Twenty-Seventh Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–10 Demcember 2013; pp. 3111–3119.
87. Pennington, J.; Socher, R.; Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Dmpirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 Octorber 2014; pp. 1532–1543.
88. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.

89. Luo, Y.; Szolovits, P.; Dighe, A.S.; Baron, J.M. 3D-MICE: integration of cross-sectional and longitudinal imputation for multi-analyte longitudinal clinical data. *J. Am. Med. Inform. Assoc.* **2017**, *25*, 645–653. [CrossRef] [PubMed]
90. Su, Y.S.; Gelman, A.; Hill, J.; Yajima, M. Multiple imputation with diagnostics (mi) in R: Opening windows into the black box. *J. Stat. Softw.* **2011**, *45*, 1–31. [CrossRef]
91. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **1977**, *39*, 1–38.
92. Chao, G.; Sun, S. Consensus and complementarity based maximum entropy discrimination for multi-view classification. *Inf. Sci.* **2016**, *367*, 296–310. [CrossRef]
93. Xu, C.; Tao, D.; Xu, C. A survey on multi-view learning. *arXiv* **2013**, arXiv:1304.5634.
94. Chao, G.; Sun, S. Alternative multiview maximum entropy discrimination. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 1445–1456. [CrossRef]
95. Chao, G.; Sun, S.; Bi, J. A survey on multi-view clustering. *arXiv* **2017**, arXiv:1712.06246.
96. Holzinger, A. From Machine Learning to Explainable AI. In Proceedings of the 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), Kosice, Slovakia, 23–25 August 2018; pp. 55–66.