

# Feature Selection and Feature Extraction in Pattern Analysis: A Literature Review

Benyamin Ghogh\*, Maria N. Samad\*, Sayema Asif Mashhadi\*,  
Tania Kapoor\*, Wahab Ali\*, Fakhri Karray, Mark Crowley

{bghogh, mnsamad, samashha, t2kapoor, wahabalikhan, karray, mcrowley}@uwaterloo.ca

Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, ON, Canada

**Abstract**—Pattern analysis often requires a pre-processing stage for extracting or selecting features in order to help the classification, prediction, or clustering stage discriminate or represent the data in a better way. The reason for this requirement is that the raw data are complex and difficult to process without extracting or selecting appropriate features beforehand. This paper reviews theory and motivation of different common methods of feature selection and extraction and introduces some of their applications. Some numerical implementations are also shown for these methods. Finally, the methods in feature selection and extraction are compared.

**Index Terms**—Pre-processing, feature selection, feature extraction, dimensionality reduction, manifold learning.

## I. INTRODUCTION

PATTERN recognition has made significant progress recently and is used for various real-world applications, from speech and image recognition to marketing and advertisement. Pattern analysis can be divided broadly into classification, regression (or prediction), and clustering, each of which tries to identify patterns in data in a specific way. The module which finds the pattern in data is named “model” in this paper. Feeding the raw data to the model, however, might not result in a satisfactory performance because the model faces a hard responsibility to find the useful information in the data. Therefore, another module is required in pattern analysis as a pre-processing stage to extract the useful information concealed in the input data. This useful information can be in terms of either better representation of data or better discrimination of classes in order to help the model perform better. There are two main categories of methods for this goal, i.e., feature selection and feature extraction.

To formally introduce feature selection and extraction, first, some notations should be defined. Suppose that there exist  $n$  data samples (points), denoted by  $\{\mathbf{x}_i\}_{i=1}^n$ , from which we want to select or extract features. Each of these data samples,  $\mathbf{x}_i$ , is a  $d$ -dimensional column vector (i.e.,  $\mathbf{x}_i \in \mathbb{R}^d$ ). The data samples  $\{\mathbf{x}_i\}_{i=1}^n$  can be represented by a matrix  $\mathbf{X} := [\mathbf{x}_1, \dots, \mathbf{x}_n] = [\mathbf{x}^1, \dots, \mathbf{x}^n]^\top \in \mathbb{R}^{d \times n}$ . In supervised cases, the target labels are denoted by  $\mathbf{t} := [t_1, \dots, t_n]^\top \in \mathbb{R}^n$ . Throughout this paper,  $x$ ,  $\mathbf{x}$ ,  $\mathbf{X}$ ,  $\mathcal{X}$ , and

$\mathcal{S}$  denote a scalar, column vector, matrix, random variable, and set, respectively. The subscript ( $x_i$ ) and superscript ( $x^j$ ) on the vector represent the  $i$ -th sample and the  $j$ -th feature, respectively. Therefore,  $x_i^j$  denotes the  $j$ -th feature of the  $i$ -th sample. Moreover, in this paper,  $\mathbf{1}$  is a vector with entries of one,  $\mathbf{1} = [1, 1, \dots, 1]^\top$ , and  $\mathbf{I}$  is the identity matrix. The  $\ell_p$ -norm is also denoted by  $\|\cdot\|_p$ .

Both feature selection and extraction reduce the dimensionality of data [1]. The goal of both feature selection and extraction is mapping  $\mathbf{x} \mapsto \mathbf{y}$  where  $\mathbf{x} \in \mathbb{R}^d$ ,  $\mathbf{y} \in \mathbb{R}^p$ , and  $p \leq d$ . In other words, the goal is to have a better representation of data with dimension  $p$ , i.e.,  $\mathbf{Y} := [\mathbf{y}_1, \dots, \mathbf{y}_n] = [\mathbf{y}^1, \dots, \mathbf{y}^n]^\top \in \mathbb{R}^{p \times n}$ . In feature selection,  $p \leq d$  and  $\{\mathbf{y}^j\}_{j=1}^p \subseteq \{\mathbf{x}^j\}_{j=1}^d$  meaning that the set of selected features is an inclusive subset of the original features. However, **feature extraction** tries to extract a completely new set of features (dimensions) from the pattern of data rather than selecting features out of the existing attributes [2]. In other words, in feature extraction, the feature set of  $\{\mathbf{y}^j\}_{j=1}^p$  is not a subset of features of  $\{\mathbf{x}^j\}_{j=1}^d$  but is a different space. In feature extraction, often  $p \ll d$  holds.

## II. FEATURE SELECTION

Feature selection [3], [4], [5] is a method of feature reduction which maps  $\mathbf{x} \in \mathbb{R}^d \mapsto \mathbf{y} \in \mathbb{R}^p$  where  $p < d$ . The reduction criterion usually either improves or maintains the accuracy or simplifies the model complexity. When there are  $d$  number of features, the total number of possible subsets are  $2^d$ . It is infeasible to enumerate the exponential number of subsets if  $d$  is large. Therefore, we need to come up with some method that works in a reasonable time. The evaluation of the subsets is based on some criterion, which can be categorized as filter or wrapper methods [3], [6], explained in the following.

### A. Filter Methods

Consider a ranking mechanism used to grade the features (variables) and the features are then removed by setting a threshold [7]. These ranking methods are categorized as filter methods because they filter the features before feeding to a learning model. Filter methods are based on two concepts “relevance” and “redundancy”, where the former is

\*The first five authors contributed equally to this work.

dependence (correlation) of feature with target and the latter addresses whether the features share redundant information. In the following, the common filter methods are introduced.

1) *Correlation Criteria*: Correlation Criteria (CC), also known as Dependence Measure (DM), is based on the relevance (predictive power) of each feature. The predictive power is computed by finding the correlation between the independent feature  $\mathbf{x}^j$  and the target (label) vector  $\mathbf{t}$ . The feature with the highest correlation value will have the highest predictive power and hence will be most useful. The features are then ranked according to some correlation-based heuristic evaluation function [8]. One of the widely used criteria for this type of measure is Pearson Correlation Coefficient (PCC) [7], [9] defined as:

$$\rho_{\mathbf{x}^j, \mathbf{t}} := \frac{\text{cov}(\mathbf{x}^j, \mathbf{t})}{\sqrt{\text{var}(\mathbf{x}^j) \text{var}(\mathbf{t})}}, \quad (1)$$

where  $\text{cov}(\cdot, \cdot)$  and  $\text{var}(\cdot)$  denote the covariance and variance, respectively.

2) *Mutual Information*: Mutual Information (MI), also known as Information Gain (IG), is the measure of dependence or shared information between two random variables [10]. It is also described as Information Theoretic Ranking Criteria (ITRC) [7], [9], [11]. The MI can be described using the concept given by Shannon's definition of entropy:

$$H(\mathbf{X}) := - \sum_{\mathbf{x}} p(\mathbf{x}) \log(p(\mathbf{x})), \quad (2)$$

which gives the uncertainty in the random variable  $\mathbf{X}$ . In feature selection, we need to maximize the mutual information (i.e., relevance) between the feature and the target variable. The mutual information (MI), which is the relative entropy between the joint distribution and product distribution, is:

$$\text{MI}(\mathbf{X}; \mathbf{T}) := \sum_{\mathbf{x}} \sum_{\mathbf{t}} p(\mathbf{x}^j, \mathbf{t}) \log \frac{p(\mathbf{x}^j, \mathbf{t})}{p(\mathbf{x}^j)p(\mathbf{t})}, \quad (3)$$

where  $p(\mathbf{x}^j, \mathbf{t})$  is the joint probability density function of feature  $\mathbf{x}^j$  and target  $\mathbf{t}$ . The  $p(\mathbf{x}^j)$  and  $p(\mathbf{t})$  are the marginal density functions. The MI is zero or greater than zero if  $\mathbf{X}$  and  $\mathbf{Y}$  are independent or dependent, respectively. For maximizing this MI, a greedy step-wise selection algorithm is adopted. In other words, there is a subset of features, denoted by matrix  $\mathcal{S}$ , initialized with one feature and features are added to this subset one by one. Suppose  $Y_{\mathcal{S}}$  denotes the matrix of data having features whose indices exist in  $\mathcal{S}$ . The index of selected feature is determined as [12]:

$$j := \arg \max_{j \notin \mathcal{S}} \text{MI}(Y_{\mathcal{S}} \cup \mathbf{x}^j; \mathbf{t}). \quad (4)$$

The above equation is also used for selecting the initial feature. Note that it is assumed the selected features are independent. The stop criterion for adding the new features is when there is highest increase in the MI at the previous step. It is noteworthy that if a variable is redundant or not informative in isolation, this does not mean that it cannot be useful when combined with another variable [13]. Moreover, this approach can reduce the dimensionality of features without having any significant negative impact on the performance [14].

3)  $\chi^2$  *Statistics*: The  $\chi^2$  Statistics method measures the dependence (relevance) of feature occurrence on the target value [14] and is based on the  $\chi^2$  probability distribution [15], defined as  $Q = \sum_{i=1}^k Z_i^2$ , where  $Z_i$ 's are independent random variables with standard normal distribution and  $k$  is the degree of freedom. This method is only applicable to cases where the target and features can take only discrete finite values. Suppose  $n_{p,q}$  denotes the number of samples which have the  $p$ -th value of the  $j$ -th feature and the  $q$ -th value of the target  $\mathbf{t}$ . Note that if the  $j$ -th feature and the target can have  $p$  and  $q$  possible values, respectively, then  $p = \{1, \dots, p\}$  and  $q = \{1, \dots, q\}$ . A contingency table is formed for each feature using the  $n_{p,q}$  for different values of target and the feature. The  $\chi^2$  measure for the  $j$ -th feature is obtained by:

$$\chi^2(\mathbf{x}^j, \mathbf{t}) := \sum_{p=1}^p \sum_{q=1}^q \frac{(n_{p,q} - \mathbb{E}_{p,q})^2}{\mathbb{E}_{p,q}}, \quad (5)$$

where  $\mathbb{E}_{p,q}$  is the expected value for  $n_{p,q}$  and is obtained as:

$$\mathbb{E}_{p,q} = n \times \Pr(p) \times \Pr(q) = n \times \frac{\sum_{q=1}^q n_{p,q}}{n} \times \frac{\sum_{p=1}^p n_{p,q}}{n}. \quad (6)$$

The  $\chi^2$  measure is calculated for all the features. The large  $\chi^2$  shows the significant dependence of the feature to the target; therefore, if it is less than a pre-defined threshold, the feature can be discarded. Note that  $\chi^2$  Statistics face a problem when some of the values of features have very low frequency. The reason for this problem is that the expected value in equation (6) becomes inaccurate at low frequencies.

4) *Markov Blanket*: Feature selection based on Markov Blanket (MB) [16] is a category of methods based on relevance. MB considers every feature  $\mathbf{x}^j$  and the target  $\mathbf{t}$  as a node in a faithful Bayesian network. The MB of a node is defined as the parents, children, and spouses of that node; therefore, in a graphical model perspective, the nodes in MB of a node suffice for estimating that node. In MB methods, the MB of a target node is found, the features in that MB are selected, and the rest are removed. Different methods have been proposed for finding the MB of target, such as Incremental Association MB (IAMB) [17], Grow-Shrink (GS) [18], Koller-Sahami (KS) [19], and Max-Min MB (MMMB) [20]. The MMMB as one of the best methods in MB is introduced here. In MMMB, first, the parents and children of target node are found. For that, those features are selected (denoted by  $\mathcal{S}$ ) that have the maximum dependence with the target, given the subset of  $\mathcal{S}$  with the minimum dependence with the target. This set of features includes some false positive nodes which are not parent or child of the target. Thus, the false positive features are filtered out if the selected features and the target are independent given any subset of selected features. Note that the  $G^2$  test, Fisher correlation, Spearman correlation, and Pearson correlation can be used for the conditional independence test. Next, spouses of target are found to form the MB with the  $\mathcal{S}$  (previously found parents and children). For this, the same procedure is done for the nodes in set  $\mathcal{S}$  to find the spouses,

grandchildren, grandparents, and siblings of the target node. Everything except the spouses should be filtered out; thus, if a feature is dependent on the target given any subset of nodes having their common child, it is retained and the rest are removed. The parents, children, and spouses of the target found form the MB.

5) *Consistency-based Filter*: Consistency-Based Filters (CBF) use a consistency measure which is based on both relevance and redundancy and is a selection criterion that aims to retain the discrimination power of the data defined by original features [21]. The InConsistency Rate (ICR) of all features is calculated via the following steps. First, inconsistent patterns are found; these are defined as patterns that are identical but are assigned to two or more different class labels, e.g., samples having patterns (0, 1) and (0, 1) in two features and with targets 0 and 1, respectively. Second, the inconsistency count for a pattern of a feature subset is calculated by taking the total number of times the inconsistent pattern occurs in the entire dataset minus the largest number of times it occurs with a certain target (label) among all targets. For example, assume  $c_1$ ,  $c_2$ , and  $c_3$  are the number of samples having targets 1, 2, and 3, respectively. If  $c_3$  is the largest, the inconsistency count will be  $c_1 + c_2$ . Third, ICR of a feature subset is the sum of inconsistency counts over all patterns (in a feature subset, there can be multiple inconsistent patterns) divided by the total number of samples  $n$ . The feature subset  $S$  with  $ICR(S) \leq \delta$  is considered to be consistent, where  $\delta$  is a pre-defined threshold. This threshold is included to tolerate noisy data. For selecting a feature subset  $S$  in CBF methods, FocusM and Automatic-Branch-and-Bound (ABB) are exhaustive search methods that can be used. This yields the smallest subset of consistent features. Las Vegas Filter (LVF), SetCover, and Quick-Branch-and-Bound (QBB) are faster search methods and can be used for large datasets when exhaustive search is not feasible.

6) *Fast Correlation-based Filter*: Fast Correlation-Based Filter (FCBF) [22] uses an entropy-based measure called Symmetrical Uncertainty (SU) to find correlation, for both relevance and redundancy, as follows:

$$SU(X, Y) := 2 \left( \frac{IG(X; Y)}{H(X) + H(Y)} \right), \quad (7)$$

where  $IG(X; Y) = H(X) - H(X|Y)$  is the information gain and  $H(X)$  is entropy defined by equation (2). The value  $SU(x^i, t)$  between a feature  $x^i$  and target  $t$  is calculated for each feature. A threshold value for SU is used to determine whether a feature is relevant or not. A subset of features  $S$  is decided by this threshold. To find out if a relevant feature  $x^i$  is redundant or not,  $SU(x^i, x^j)$  value between two features  $x^i$  and  $x^j$  is calculated. For a feature  $x^i$ , all its redundant features are collected together (denoted by  $S_{P_i}$ ) and divided in two subsets,  $S_{P_i}^+$  and  $S_{P_i}^-$ , where  $S_{P_i}^+ = \{x^j | x^j \in S_{P_i}, SU(x^j, t) > SU(x^i, t)\}$  and  $S_{P_i}^- = \{x^j | x^j \in S_{P_i}, SU(x^j, t) \leq SU(x^i, t)\}$ . All features in  $S_{P_i}^+$  are processed before making a decision on  $x^i$ . If a feature is predominant (i.e., its  $S_{P_i}^+$  is empty) then all features in  $S_{P_i}^-$  are removed and  $x^i$  is retained. The feature with the

largest  $SU(x^i, t)$  value is a predominant feature and used as a starting point to eliminate other features. FCBF uses the novel idea of predominant correlation to make feature selection faster and more efficient so that it can be easily used on high dimensional data. This method greatly reduces dimensionality and increases classification accuracy.

7) *Interact*: In feature selection, many algorithms apply correlation metrics to find which feature correlates most to the target. These algorithms single out features and do not consider the combined effect of two or more features with the target. In other words, some features might not have individual effect but alongside other features they give high correlation to the target and increase classification performance. Interact [23] is a fast filter algorithm that searches for interacting features. First, Symmetrical Uncertainty (SU), defined by equation (7), is used to evaluate the correlation of individual features with the target. This heuristic is used to rank features in descending order such that the most important feature is positioned at the beginning of the list  $S$ . Interact also uses Consistency Contribution (cc) or c-contribution metric which is defined as:

$$cc(x^i, X) := ICR(X \setminus x^i) - ICR(X), \quad (8)$$

where  $x^i$  is the feature for which cc is being calculated, and ICR stands for inconsistency rate defined in Section II-A.5. The “ $\setminus$ ” is the set-theoretic difference operator, i.e.,  $X \setminus x^i$  means  $X$  excluding the feature  $x^i$ . The cc is calculated for each feature from the end of the list  $S$  and if the cc of a feature is less than a pre-defined threshold, that feature is eliminated. This backward elimination makes sure that the cc is first calculated for the features having small correlation with the target. The appropriate pre-defined threshold is assigned based on cross validation. The Interact method has the advantage of being fast.

8) *Minimal-Redundancy-Maximal-Relevance*: The Minimal Redundancy Maximal Relevance (mRMR) [24], [25] is based on maximizing the relevance and minimizing redundancy of features. The relevance of features means that each of the selected features should have the largest relevance (correlation or mutual information [26]) with the target  $t$  for having better discrimination [11]. This dependence or relevance is formulated as the average mutual information between the selected features (set  $S$ ) [25]:

$$d := \frac{1}{|S|} \sum_{x^i \in S} MI(x^i; t), \quad (9)$$

where MI is the mutual information defined by equation (3). The redundancy of features, on the other hand, should be minimized because having at least one of the redundant features suffices for a good performance. The redundancy of features  $x^i$  and  $x^j$  is formulated as [25]:

$$r := \frac{1}{|S|^2} \sum_{x^i, x^j \in S} MI(x^i; x^j). \quad (10)$$

By defining  $\phi = d - r$ , the goal of mRMR is to maximize  $\phi$  [25]. Therefore, the features which maximize  $\phi$  are added to the set  $S$  incrementally and one by one [25].

## B. Wrapper Methods

As seen previously, filter methods select the optimal features to be passed to the learning model, i.e., classifier, regression, etc. Wrapper methods, on the other hand, integrate the model within the feature subset search. In this way, different subsets of features are found or generated and evaluated through the model. The fitness of a feature subset is evaluated by training and testing it on the model. Thus in this sense, the algorithm for the search of the best suboptimal subset of the feature set is essentially “wrapped” around the model. The search for the best subset of the feature set, however, is an NP-hard problem. Therefore, heuristic search methods are used to guide the search. These search methods can be divided in two categories: Sequential and Metaheuristic algorithms.

1) *Sequential Selection Methods*: Sequential feature selection algorithms access the features from the given feature space in a sequential manner. These algorithms are called sequential due to the iterative nature of the algorithms. There are two main categories of sequential selection methods: Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) [27]. Although both the methods switch between including and excluding features, they are based on two different algorithms according to the dominant direction of the search. In SFS, the set of selected features, denoted by  $\mathcal{S}$ , is initially empty. The features are added to  $\mathcal{S}$  one by one if they improve the model performance the best. This procedure is repeated until the required number of features are selected. In SBS, on the other hand, the set  $\mathcal{S}$  is initialized by the entire features and features are removed sequentially based on the performance [3]. The SFS and SBS ignore the dependence of features, i.e., some specific features might have better performance than the case where those features are used alongside some other features. Therefore, Sequential Floating Forward Selection (SFFS) and Sequential Floating Backward Selection (SFBS) [28] were proposed. In SFFS, after adding a feature as in SFS, every feature is tested for being excluded if the performance improves. Similar approach is performed in SFBS but with opposite direction.

2) *Metaheuristic Methods*: The metaheuristic algorithms are also referred to as evolutionary algorithms. These methods have low implementation complexity and can adapt to a variety of problems. They are also less prone to get stuck in a local optima as compared to sequential methods, where the objective function is the performance of the model. Many metaheuristic methods have been used for feature selection such as the binary dragonfly algorithm used in [29]. Another recent technique called whale optimization algorithm was used for feature selection and explored against other common metaheuristic methods in [30]. As examples of metaheuristic approaches for feature selection, the two broad metaheuristic methods, Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), are explained here. PSO is inspired by the social behavior observed in some animals like the flocking of birds or formation of schools of fish and GA is inspired by natural selection and has been used widely in search and

optimization problems.

The original version of PSO [31] was developed for continuous optimization problems whose potential solutions are represented by particles. However, many problems are combinatorial optimization problems usually with binary decisions. Geometric PSO (GPSO) uses a geometric framework described in [32] to provide a version of PSO that can be generalized to any solution representation. GPSO considers the current position of particle, the global best position, and the local best position of the particle as the three parents of particle and creates the offspring (similar to GA approach) by applying a three-Parent Mask-Based Crossover (3PMBCX) operator on them. The 3PMBCX simply takes each element of crossover with specific probabilities from the three parents. The GPSO is used in [33] for the sake of feature selection. Binary PSO [34] can also be used for feature selection where binary particles represent selection of features.

In GA, the potential solutions are represented by chromosomes [35]. For feature selection, the genes in the chromosome correspond to features and can take values 1 or 0 for selection or not selection of feature, respectively. The generations of chromosomes improve by crossovers and mutations until convergence. The GA is used in different works such as [36], [37] for selecting features. A modified version of GA, named CHCGA [38], is used in [39] for feature selection. The CHCGA maintains diversity and avoids stagnation of the population by using a Half Uniform Crossover (HUX) operator which crosses over half of the non-matching genes at random. One of the problems of GA is poor initialization of chromosomes. A modified version of GA proposed in [40] for feature selection reduces the risk of poor initialization by introducing some excellent chromosomes as well as random chromosomes for initialization. In that work, the estimated regression coefficients  $\hat{\beta}_1, \dots, \hat{\beta}_d$  and their estimated standard deviations  $\sigma_{\hat{\beta}_1}, \dots, \sigma_{\hat{\beta}_d}$  are calculated using regression, then Student's  $t$ -test is applied for every coefficient  $t_j \leftarrow (\hat{\beta}_j - 0) / \sigma_{\hat{\beta}_j}$ . The binary excellent chromosome  $\mathbf{s} = [s_1, \dots, s_d]^T$  is generated where  $s_j = 1$  with probability  $|t_j| / \sum_{i=1}^d |t_i|$ .

## III. FEATURE EXTRACTION

As features of data are not necessarily uncorrelated (matrix  $\mathbf{X}$  is not full rank), they share some information and thus there usually exists dummy information in the data pattern. Hence, in mapping  $\mathbf{x} \in \mathbb{R}^d \mapsto \mathbf{y} \in \mathbb{R}^p$  usually  $p \ll d$  or at least  $p < d$  holds, where  $p$  is named the intrinsic dimension of data [41]. This is also referred to as the manifold hypothesis [42] stating that the data points exist on a lower dimensional sub-manifold or subspace. This is the reason that feature extraction and dimensionality reduction are often used interchangeably in the literature. The main goal of dimensionality reduction is usually either better representation/discrimination of data [43] or easier visualization of data [44], [45]. It is noteworthy that the  $\mathbb{R}^p$  space is referred to as feature space (i.e., *feature extraction*), embedded space (i.e., *embedding*), encoded space

(i.e., *encoding*), subspace (i.e., *subspace learning*), lower-dimensional space (i.e., *dimensionality reduction*) [46], sub-manifold (i.e., *manifold learning*) [47], or representation space (i.e., *representation learning*) [48] in the literature.

The dimensionality reduction methods can be divided into two main categories, i.e., supervised and unsupervised methods. Supervised methods take into account the labels and classes of data samples while the unsupervised methods are based on the variation and pattern of data. Another categorization of feature extraction is dividing methods into linear and non-linear. The former assumes that the data falls on a linear subspace or classes of data can be distinguished linearly, while the latter supposes that the pattern of data is more complex and exists on a non-linear sub-manifold. In the following, we explore the different methods of feature extraction. For the sake of brevity, we mention the basics and forgo some very detailed methods and improvements of these methods such as using Nyström Method for incorporating out-of-sample data [49].

#### A. Unsupervised Feature Extraction

Unsupervised methods in feature extraction do not use the labels/classes of data for feature extraction [50], [51]. In other words, they do not respect the discrimination of classes but they mostly concentrate on the variation and distribution of data.

1) *Principal Component Analysis*: Principal Component Analysis (PCA) [52], [53] was first proposed by [54]. As a linear unsupervised method, it tries to find the directions which represent the variation of data the best. The original coordinates do not necessarily represent the direction of variation. The aim of PCA is to find orthogonal directions which represent the data with the least error [55]. Therefore, PCA can be also considered as rotating the coordinate system [56].

The projection of data onto direction  $\mathbf{u}$  is  $\mathbf{u}^\top \mathbf{X}$ . Taking  $\mathbf{S}$  to be the  $d \times d$  covariance matrix of data, the variance of this projection is  $\mathbf{u}^\top \mathbf{S} \mathbf{u}$ . PCA tries to maximize this variance to find the most variant orthonormal directions of data. Solving this optimization problem [55] yields to  $\mathbf{S} \mathbf{u} = \lambda \mathbf{u}$  which is the eigenvalue problem [57] for the covariance matrix  $\mathbf{S}$ . Therefore, the desired directions (columns of matrix  $\mathbf{U}$ ) are the eigenvectors of the covariance matrix of data. The eigenvectors, sorted by their eigenvalues in descending order, represent the largest to smallest variations of data and are named principal directions or axes. The features (rows) of the projected data  $\mathbf{U}^\top \mathbf{X}$  are named principal components. Usually, the components with smallest eigenvalues are cut off to reduce the data. There are different methods for estimating the best number of components to keep (denoted by  $p$ ), such as using Bayesian model selection [58], scree plot [59], and comparing the ratio  $\lambda_i / \sum_{j=1}^d \lambda_j$  with a threshold [53] where  $\lambda_i$  denotes the eigenvalue related to the  $i$ -th principal component.

In this paper, out-of-sample data refers to the data sample which does not exist in the training set of samples from which the subspace was created. Assume  $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_d] \in$

$\mathbb{R}^{p \times d}$ . The projection of training data  $\mathbf{X}$  and the out-of-sample data  $\mathbf{x}$  are  $\mathbf{Y} = \mathbf{U}^\top \mathbf{X} \in \mathbb{R}^{p \times n}$  and  $\mathbf{y} = \mathbf{u}^\top \mathbf{x}$ , respectively. Note that the projected data can also be reconstructed back but it will have distortion error because the samples were projected onto a subspace. The reconstruction of training and out-of-sample data are  $\hat{\mathbf{X}} = \mathbf{U} \mathbf{Y}$  and  $\hat{\mathbf{x}} = \mathbf{U} \mathbf{y}$ , respectively.

2) *Dual Principal Component Analysis*: The explained PCA was based on eigenvalue decomposition; however, it can be done based on Singular-Value Decomposition (SVD) more easily [55]. Considering  $\mathbf{X} = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$ , the  $\mathbf{U}$  is exactly the same as before and contains the eigenvectors (principal directions) of  $\mathbf{X} \mathbf{X}^\top$  (the covariance matrix). The matrix  $\mathbf{V}$  contains the eigenvectors of  $\mathbf{X}^\top \mathbf{X}$ . In cases where  $d \gg n$ , such as in images, calculating eigenvectors of  $\mathbf{X} \mathbf{X}^\top$  with size  $d \times d$  is less efficient than  $\mathbf{X}^\top \mathbf{X}$  with size  $n \times n$ . Hence, in these cases, dual PCA is used rather than the ordinary (or direct) PCA [55]. In dual PCA, projection and reconstruction of training data ( $\mathbf{X}$ ) and out-of-sample data ( $\mathbf{x}$ ) are formulated as:

$$\text{Projection for } \mathbf{X} : \quad \mathbf{Y} = \mathbf{\Sigma} \mathbf{V}^\top, \quad (11)$$

$$\text{Projection for } \mathbf{x} : \quad \mathbf{y} = \mathbf{\Sigma}^{-1} \mathbf{V}^\top \mathbf{X}^\top \mathbf{x}, \quad (12)$$

$$\text{Reconstruction for } \mathbf{X} : \quad \hat{\mathbf{X}} = \mathbf{X} \mathbf{V} \mathbf{V}^\top, \quad (13)$$

$$\text{Reconstruction for } \mathbf{x} : \quad \hat{\mathbf{x}} = \mathbf{X} \mathbf{V} \mathbf{\Sigma}^{-2} \mathbf{V}^\top \mathbf{X}^\top \mathbf{x}. \quad (14)$$

3) *Kernel Principal Component Analysis*: PCA tries to find the linear subspace for representing the pattern of data. However, kernel PCA [60] finds the non-linear subspace of data which is useful if the data pattern is not linear. The kernel PCA uses kernel method which maps data to a higher dimensional space  $\mathbf{x} \mapsto \phi(\mathbf{x})$  where  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$  and  $d' > d$  [61], [62]. Note that having high dimensions has both its blessings and curses [63]. The “curse of dimensionality” refers to the fact that by going to higher dimensions, the number of samples required for learning a function grows exponentially. On the other hand, the “blessing of dimensionality” states that in higher dimensions, the representation or discrimination of data is easier. Kernel PCA relies on the blessing of dimensionality by using kernels.

The kernel matrix is  $K(\mathbf{x}_1, \mathbf{x}_2) = \phi(\mathbf{x}_1)^\top \phi(\mathbf{x}_2)$  which replaces  $\mathbf{x}_1^\top \mathbf{x}_2$  using the kernel trick. The most popular kernels are linear kernel  $\mathbf{x}_1^\top \mathbf{x}_2$ , polynomial kernel  $(1 + \langle \mathbf{x}_1, \mathbf{x}_2 \rangle)^i$ , and Gaussian kernel  $\exp(-\|\mathbf{x}_1, \mathbf{x}_2\|_2^2 / 2\sigma^2)$ , where  $i$  is a positive integer and denotes the polynomial grade of kernel. Note that Gaussian kernel is also referred to as Radial Basis Function (RBF) kernel.

After the kernel trick, PCA is applied on the data. Therefore, in kernel PCA, SVD is applied on the kernel matrix  $K(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{U} \mathbf{\Sigma} \mathbf{V}^\top$  rather than on  $\mathbf{X}$ . The projections of training and out-of-sample data are formulated as:

$$\text{Projection for } \mathbf{X} : \quad \mathbf{Y} = \mathbf{\Sigma} \mathbf{V}^\top, \quad (15)$$

$$\text{Projection for } \mathbf{x} : \quad \mathbf{y} = \mathbf{\Sigma}^{-1} \mathbf{V}^\top K(\mathbf{X}, \mathbf{x}). \quad (16)$$

Note that reconstruction cannot be done in kernel PCA [55]. It is noteworthy that kernel PCA usually does not perform satisfactorily in practice [55], [64] and the reason of it is

the unknown perfect choice of kernels. However, it provides technical support for other methods which are explained later.

4) *Multidimensional Scaling*: Multidimensional Scaling (MDS) [65] is a method for dimensionality reduction and feature extraction. It includes two main approaches, i.e., metric (classic) and non-metric. We cover the classic approach here. The metric MDS is also called Principal Coordinate Analysis (PCoA) [66] in the literature. The goal of MDS is to preserve the pairwise Euclidean distance or the similarity (inner product) of data samples in the feature space. The solution to this goal [55] is:

$$\mathbf{Y} = \hat{\mathbf{\Lambda}}^{\frac{1}{2}} \mathbf{V}^T, \quad (17)$$

where  $\hat{\mathbf{\Lambda}} \in \mathbb{R}^{p \times p}$  is a diagonal matrix having the top  $p$  eigenvalues of  $\mathbf{X}^T \mathbf{X}$  and  $\mathbf{V} \in \mathbb{R}^{n \times p}$  contains the eigenvectors of  $\mathbf{X}^T \mathbf{X}$  corresponding to the top  $p$  eigenvalues. Comparing equations (11) and (17) shows that if the distance metric in MDS is Euclidean distance, the solutions of metric MDS and dual PCA are identical. Note that what MDS does is converting distance matrix of samples, denoted by  $\mathbf{D}^{(\mathbf{X})}$ , to a kernel matrix  $\mathbf{K}$  [55] formulated as:

$$\mathbf{K} = -\frac{1}{2} \mathbf{H} \mathbf{D}^{(\mathbf{X})} \mathbf{H}, \quad (18)$$

where  $\mathbf{H} := \mathbf{I} - \frac{1}{n} \mathbf{1} \mathbf{1}^T \in \mathbb{R}^{n \times n}$  is the centering matrix used for double-centering the distance matrix. When  $\mathbf{D}^{(\mathbf{X})}$  is based on Euclidean distance, the kernel  $\mathbf{K}$  is positive semi-definite and is equal to  $\mathbf{K} = \mathbf{X}^T \mathbf{X}$  [65]. Therefore,  $\hat{\mathbf{\Lambda}}$  and  $\mathbf{V}$  are the eigenvalues and eigenvectors of the kernel matrix, respectively.

5) *Isomap*: Linear methods, such as PCA and MDA (with Euclidean distance), have the lack of not capturing the possible non-linear essence of pattern. For example, suppose the data exist on a non-linear manifold. When applying PCA, the samples on the different sides of manifold mistakenly fall next to each other because PCA cannot capture the structure of the non-linear manifold [56]. Therefore, other methods are required which consider the distances of samples on the manifold. Isomap [67] is a method which considers the geodesic distance of data samples on the manifold. For approximating the geodesic distances, it firstly constructs a  $k$ -nearest neighbor graph on the  $n$  data samples. Then it computes the shortest path distances between all pairs of samples resulting in the geodesic distance matrix  $\mathbf{D}^{(G)}$ . Different algorithms can be used for finding the shortest paths, such as the Dijkstra and Floyd-Warshall algorithms [68]. Finally, it runs the metric MDS using the kernel based on  $\mathbf{D}^{(G)}$  as:

$$\mathbf{K} = -\frac{1}{2} \mathbf{H} \mathbf{D}^{(G)} \mathbf{H}. \quad (19)$$

The embedded data points are obtained from equation (17) but with  $\hat{\mathbf{\Lambda}}$  and  $\mathbf{V}$  as the eigenvalues and eigenvectors of equation (19), respectively.

6) *Locally Linear Embedding*: Another perspective toward capturing the non-linear manifold of data pattern is to consider the manifold as integration of small linear patches. This intuition is similar to piece-wise linear (spline) regression [69]. In other words, unfolding the manifold can

be approximated by locally capturing the piece-wise linear patches and putting them together. For this goal, Locally Linear Embedding (LLE) [70], [71] is proposed which first constructs a  $k$ -nearest neighbor graph similar to Isomap. Then it tries to locally represent every data sample  $\mathbf{x}_i$  using a weighted summation of its  $k$ -nearest neighbors. Taking  $\mathbf{w}_i$  as the  $i$ -th row entry of the  $n \times k$  weight matrix  $\mathbf{W}$ , the solution to this goal is:

$$\mathbf{w}_i = \frac{1}{\mathbf{1}^T \mathbf{G}_i^{-1} \mathbf{1}} \mathbf{G}_i^{-1} \mathbf{1}, \quad (20)$$

$$\mathbf{G} := (\mathbf{x}_i \mathbf{1}^T - \mathbf{V}_i)^T (\mathbf{x}_i \mathbf{1}^T - \mathbf{V}_i), \quad (21)$$

where  $\mathbf{G}$  is named Gram matrix and  $\mathbf{V}$  is a  $d \times k$  matrix defined as  $\mathbf{V} := [\mathbf{x}_{v_{i(1)}}, \dots, \mathbf{x}_{v_{i(k)}}]$ . The  $v_{i(j)}$  denotes the index of the  $j$ -th neighbor of sample  $\mathbf{x}_i$  among the  $n$  samples. Note that dividing by  $\mathbf{1}^T \mathbf{G}_i^{-1} \mathbf{1}$  in equation (20) is for normalizing weights associated with the  $i$ -th sample so that  $\sum_{j=1}^k w_{ij} = 1$  is satisfied, where  $w_{ij}$  is the entry of  $\mathbf{W}$  in the  $i$ -th row and the  $j$ -th column.

After representing the samples as a weighted summation of their neighbors, LLE tries to represent the samples in the lower dimensional space (denoted by  $\mathbf{y}$ ) by their neighbors with the same obtained weights. In other words, it preserves the locality of data pattern in the feature space. The solution to this second goal (matrix  $\mathbf{Y}$ ) is the  $p$  smallest eigenvectors of  $\mathbf{M}$  after ignoring an eigenvector having eigenvalue equal to zero. The  $n \times n$  matrix  $\mathbf{M}$  is

$$\mathbf{M} := (\mathbf{I} - \mathbf{W}')(\mathbf{I} - \mathbf{W}')^T, \quad (22)$$

where  $\mathbf{W}'$  is considered as a  $n \times n$  weight matrix having zero entries for the non-neighbor samples.

7) *Laplacian Eigenmap*: Another non-linear perspective to dimensionality reduction is to preserve locality based on the similarity of neighbor samples. Laplacian Eigenmap (LE) [72] is a method which tracks this aim. This method, first, constructs a weighted graph  $\mathbf{G}$  in which vertices are data samples and edge weights demonstrate a measure of similarity such as  $w_{ij} = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / \gamma)$ . LE tries to capture the locality. It minimizes  $\|\mathbf{y}_i - \mathbf{y}_j\|_2^2$  when the samples  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are close to each other, i.e., weight  $w_{ij}$  is large. The solution to this problem [55] is the  $p$  smallest eigenvectors of the Laplacian matrix of graph  $\mathbf{G}$  defined as  $\mathbf{L} := \mathbf{D} - \mathbf{W}$ , where  $\mathbf{D}$  is a diagonal matrix with elements  $d_{ii} = \sum_j w_{ij}$ . Note that according to characteristics of Laplacian matrix, for the connected graph  $\mathbf{G}$ , there exists one zero eigenvalue whose corresponding eigenvector should be ignored.

8) *Maximum Variance Unfolding*: Surprisingly, all the unsupervised methods explained so far can be seen as the kernel PCA with different kernels [73], [74]:

$$\begin{cases} \text{PCA:} & \mathbf{K} = \mathbf{X}^T \mathbf{X}, \\ \text{MDS:} & \mathbf{K} = -\frac{1}{2} \mathbf{H} \mathbf{D}^{(\mathbf{X})} \mathbf{H}, \\ \text{Isomap:} & \mathbf{K} = -\frac{1}{2} \mathbf{H} \mathbf{D}^{(G)} \mathbf{H}, \\ \text{LLE:} & \mathbf{K} = \mathbf{M}^\dagger, \\ \text{LE:} & \mathbf{K} = \mathbf{L}^\dagger, \end{cases}$$

where  $M^\dagger$  and  $L^\dagger$  are the pseudo-inverses of matrices  $M$  and  $L$ , respectively. The reason of this pseudo-inverse is that in LLE and LE, the eigenvectors having smallest, rather than the largest, eigenvalues are considered. Inspired by this idea, another approach toward dimensionality reduction is to apply kernel PCA but with the optimum kernel. In other words, the best kernel can be found using optimization. This is the aim of Maximum Variance Unfolding (MVU) [75], [76], [77]. The reason for this name is that it finds the best kernel which maximizes the variance of data in the feature space. The variance of data is equal to the trace of kernel matrix, denoted by  $\text{tr}(\mathbf{K})$ , which is equal to the summation of eigenvalues of  $\mathbf{K}$  (recall that we saw in PCA that eigenvalues show the amount of variance). Supposing that  $\mathbf{K}_{ij}$  denotes the  $(i, j)$  entry of the kernel matrix, the optimization problem which MVU tackles is:

$$\begin{aligned} & \underset{\mathbf{K}}{\text{maximize}} \quad \text{tr}(\mathbf{K}) \\ & \text{subject to} \quad \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 = \mathbf{K}_{ii} + \mathbf{K}_{jj} - 2\mathbf{K}_{ij}, \\ & \quad \sum_{i=1}^n \sum_{j=1}^n \mathbf{K}_{ij} = 0, \\ & \quad \mathbf{K} \succeq 0, \end{aligned} \quad (23)$$

which is a semi-definite programming optimization problem [78]. That is why MVU is also addressed as Semi-definite Embedding (SDE) [76]. This optimization problem does not have a closed form solution and needs optimization toolboxes to be solved [75].

9) *Autoencoders & Neural Networks*: Autoencoders (AEs), as neural networks, can be used for compression, feature extraction or, in general, for data representation [79], [80]. The most basic form of an AE is with an encoder and decoder having just one hidden layer. The input is fed to the encoder and output is extracted from the decoder. The output is the reconstruction of the input; therefore, the number of nodes of the encoder and decoder are the same. The hidden layer usually has fewer number of nodes than the encoder/decoder layer. The AEs with less or more number of hidden neurons are called undercomplete and overcomplete, respectively [81]. AEs were first introduced in [82].

AEs try to minimize the error between input  $\{\mathbf{x}_i\}_{i=1}^n$  and decoded output  $\{\hat{\mathbf{x}}_i\}_{i=1}^n$ , i.e., reproduce the exact input using the embedded information in the hidden layer. The hidden layer in undercomplete AE is the representation of data with reduced dimensionality and compressed form [80]. Once the network is trained, the decoder part is removed and output of the innermost hidden layer is used for feature extraction from input. To get better compression or greater dimensionality reduction, multiple hidden layers should be used resulting in deep AE [81]. Previously, training deep AE faced the problem of vanishing gradients because of large number of layers. This problem was first resolved by considering each pair of layers as a Restricted Boltzmann Machine (RBM) and training it in unsupervised manner [80]. This AE is also referred to as Deep Belief Network (DBN) [83]. The obtained weights are then fine tuned by backpropagation. However, recently, vanishing gradients is resolved mostly

because of using ReLu activation function [84] and batch normalization [85]. Therefore, the current learning algorithm used in AEs is the backpropagation algorithm [86], where error is between  $\mathbf{x}_i$  and  $\hat{\mathbf{x}}_i$ .

10) *t-distributed Stochastic Neighbor Embedding*: The  $t$ -distributed Stochastic Neighbor Embedding ( $t$ -SNE) [87], which is an improvement to Stochastic Neighbor Embedding (SNE) [88], is a state-of-the-art method for data visualization and its goal is to preserve the joint distribution of data samples in the original and embedding spaces. If  $p_{ij}$  and  $q_{ij}$ , respectively, denote the probability that  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are neighbors (similar) and  $\mathbf{y}_i$  and  $\mathbf{y}_j$  are neighbors, we have:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}, \quad (24)$$

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|_2^2 / 2\sigma_i^2)}, \quad (25)$$

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|_2^2)^{-1}}{\sum_{k \neq l} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|_2^2)^{-1}}, \quad (26)$$

where  $\sigma_i^2$  is the variance of Gaussian distribution over  $\mathbf{x}_i$  obtained by binary search [88]. The  $t$ -SNE considers Gaussian and Student's  $t$ -distribution [89] for original and embedding spaces, respectively. The embedded samples are obtained using gradient descent over minimizing the Kullback-Leibler divergence [90] of  $p$  and  $q$  distributions (equations (24) and (26)). The heavy tails of  $t$ -distribution gives  $t$ -SNE the ability to deal with the problem of visualizing "crowded" high-dimensional data in a low dimensional (e.g., 2D or 3D) space [87], [91].

## B. Supervised Feature Extraction

1) *Fisher Linear Discriminant Analysis*: Fisher Linear Discriminant Analysis (FLDA) is also referred to as Fisher Discriminant Analysis (FDA) or Linear Discriminant Analysis (LDA) in literature. The base of this method was proposed by a genius named Ronald A. Fisher [92]. Similar to PCA, FLDA calculates the projection of data along a direction; however, rather than maximizing the variation of data, FLDA utilizes label information to get a projection maximizing the ratio of between-class variance to within-class variance. The goal of FLDA is formulated as the Fisher criterion [93], [94]:

$$J(\mathbf{u}) := \frac{\mathbf{u}^\top \mathbf{S}_B \mathbf{u}}{\mathbf{u}^\top \mathbf{S}_W \mathbf{u}}, \quad (27)$$

where  $\mathbf{u}$  is the projection direction, and  $\mathbf{S}_B$  and  $\mathbf{S}_W$  are between- and within-class scatters formulated as [93]:

$$\mathbf{S}_B := \sum_{c=1}^{|\mathcal{C}|} n_c (\bar{\mathbf{x}}_c - \bar{\mathbf{x}})(\bar{\mathbf{x}}_c - \bar{\mathbf{x}})^\top, \quad (28)$$

$$\mathbf{S}_W := \sum_{c=1}^{|\mathcal{C}|} \sum_{\mathbf{x}_i \in c} (\mathbf{x}_i - \bar{\mathbf{x}}_c)(\mathbf{x}_i - \bar{\mathbf{x}}_c)^\top, \quad (29)$$

assuming that  $|\mathcal{C}|$  is the number of classes,  $n_c$  is the number of training samples in class  $c$ ,  $\bar{\mathbf{x}}_c$  is the mean of class  $c$ , and  $\bar{\mathbf{x}}$  is the mean of all training samples. Maximizing the Fisher criterion  $J(\mathbf{u})$  results in a generalized eigenvalue problem  $\mathbf{S}_B \mathbf{u} = \lambda \mathbf{S}_W \mathbf{u}$  [57]. Therefore, the projection directions

(columns of the projection matrix  $U$ ) are the  $p$  eigenvectors of  $S_W^{-1}S_B$  with the largest eigenvalues. Note that as  $S_B$  has rank  $\leq (C-1)$ , we have  $p \leq C-1$  [95]. It is noteworthy that when  $n \ll d$  (e.g., in images), the  $S_W$  might become singular and not invertible. In these cases, FLDA is difficult to be directly implemented and can be applied on PCA-transformed space [96]. It is also noteworthy that an ensemble of FLDA models, named Fisher forest [97], can be useful for classifying data with different essences and even different dimensionality.

2) *Kernel Fisher Linear Discriminant Analysis*: FLDA tries to find a linear discriminant but a linear discriminant may not be enough to separate different classes in some cases. Similar to kernel PCA, the kernel trick is applied and inner products are replaced by kernel function  $K(x_1, x_2) = \phi(x_1)^\top \phi(x_2)$  [167]. In kernel FLDA, the objective function to be maximized [167] is:

$$J(u) := \frac{u^\top M u}{u^\top N u}, \quad (30)$$

where:

$$M := \sum_{c=1}^{|C|} n_c (m_c - m_*) (m_c - m_*)^\top \in \mathbb{R}^{n \times n}, \quad (31)$$

$$i\text{-th entry of } m_c (\in \mathbb{R}^n) := \frac{1}{n} \sum_{j=1}^{n_c} K(x_i, x_{j,c}), \quad (32)$$

$$i\text{-th entry of } m_* (\in \mathbb{R}^n) := \frac{1}{n} \sum_{j=1}^n K(x_i, x_j), \quad (33)$$

$$N := \sum_{c=1}^{|C|} K_c (I - \frac{1}{n_c} \mathbf{1}\mathbf{1}^\top) K_c^\top \in \mathbb{R}^{n \times n}, \quad (34)$$

$$\text{entry } (i, j) \text{ of } K_c (\in \mathbb{R}^{n \times n_c}) := K(x_i, x_{j,c}), \quad (35)$$

where  $x_{j,c}$  denotes the  $j$ -th sample in class  $c$ . Similar to the approach of FLDA, the projection directions (columns of  $U$ ) are the  $p$  ( $\leq C-1$ ) eigenvectors of  $N^{-1}M$  with the largest eigenvalues [167]. The projection of data  $X_t$  is obtained by  $U^\top K(X, X_t)$ .

3) *Supervised Principal Component Analysis*: There are various approaches that have been used to do Supervised Principal Component Analysis, such as the original method of Bair's Supervised Principal Components (SPC) [168], [169], Hilbert-Schmidt Component Analysis (HSCA) [170], and Supervised Principal Component Analysis (SPCA) [171]. Here, SPCA [171] is presented.

The Hilbert-Schmidt Independence Criterion (HSIC) [172] is a measure of dependence between two random variables. The SPCA uses this criterion for maximizing the dependence between the transformed data  $U^\top X$  and the targets  $T := [t_1, \dots, t_n]$ . Assuming that  $K = X^\top U U^\top X$  is the linear kernel over  $U^\top X$  and  $B$  is an arbitrary kernel over  $T$ , we have:

$$\text{HSIC} := \frac{1}{(n-1)^2} \text{tr}(K H B H), \quad (36)$$

where  $H = I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top$  is the centering matrix. Maximizing this HSIC criterion [171] results in the solution of  $U$  which

contains the eigenvectors, having the top  $p$  eigenvalues, of  $X H B H X^\top$ . The encoding of data  $X$  is obtained by  $U^\top X$  and its reconstruction can be done by  $\hat{X} = UY$ .

4) *Metric Learning*: The performance of many machine learning algorithms depend critically on the existence of a good distance measure (metric) over an input space, including both supervised and unsupervised learning [173]. Metric Learning (ML) is the task of learning the best distance function (distance metric) directly from the training data. It is not just one algorithm but a class of algorithms. The general form of metric [174] is usually defined as a form similar to Mahalanobis distance:

$$\|x_i - x_j\|_A := (x_i - x_j)^\top A (x_i - x_j), \quad (37)$$

where  $A = U U^\top \succeq 0$  to have a valid distance metric. Most of the Metric Learning algorithms are optimization problems where  $A$  is unknown to make data points in same class (similar pairs) closer to each other, and points in different classes far apart from each other [56]. It is easily observed that  $\|x_i - x_j\|_A = (U^\top x_i - U^\top x_j)^\top (U^\top x_i - U^\top x_j)$ , so this metric is equivalent to projection of data with projection matrix  $U$  and then using Euclidean distance in the embedded space [174]. Therefore, Metric learning can be considered as a feature extraction method [175], [176]. The first work in ML was [173]. Another popular ML algorithm is Maximally Collapsing Metric Learning (MCML) [176] which deals with probability of similarity of points. Here, metric learning with class-equivalence side information [177] is explained which has a closed-form solution. Assuming  $\mathcal{S}$  and  $\mathcal{D}$ , respectively, denote the sets of similar and dissimilar samples (regarding the targets),  $M_S$  is defined as:

$$M_S := \frac{1}{|\mathcal{S}|} \sum_{(x_i, x_j) \in \mathcal{S}} (x_i - x_j)(x_i - x_j)^\top, \quad (38)$$

and  $M_D$  is defined similarly for points in set  $\mathcal{D}$ . By minimizing the distance of similar points and maximizing the distance of dissimilar ones, it can be shown [177] that the best  $U$  in matrix  $A$  is the  $p$  eigenvectors of  $(M_S - M_D)$  having the smallest eigenvalues.

#### IV. APPLICATIONS OF FEATURE SELECTION AND EXTRACTION

There exist various applications in the literature which have used the feature selection and extraction methods. Table I summarizes some examples of these applications for the different methods introduced in this paper. As can be seen in this table, the feature selection and extraction methods are useful for different applications such as face recognition, action recognition, gesture recognition, speech recognition, medical imaging, biomedical engineering, marketing, wireless network, gene expression, software fault detection, internet traffic prediction, etc. The variety of applications of feature selection and extraction show their usefulness and effectiveness in different real-world problems.



TABLE I: Some examples of applications of different methods in feature selection and extraction.

		Method	Applications
Feature Selection	Filters	CC	network intrusion detection [98]
		MI	advertisement [99], action recognition [100]
		$\chi^2$ Statistics	medical imaging [101], text classification [102], [103]
		MB	Gaussian mixture clustering [104]
		CBF	IP traffic [105], credit scoring [106], fault detection [107], antidepressant medication [108]
		FCBF	software defect prediction [109], internet traffic [110], intelligent tutoring [111], network fault diagnosis [112]
		Interact	network intrusion detection [113], automatic recommendation [114], dry eye detection [115]
		mRMR	health monitoring [116], churn prediction [117], gene expression [118]
	Wrappers	SS	satellite images [119], medical imaging [120]
		Metaheuristic	cancer classification [121], hospital demands [40]
Feature Extraction	Unsupervised	PCA	face recognition [122], action recognition [123], EEG [124], object orientation detection [125]
		Kernel PCA	face recognition [126], [127], fault detection [128]
		MDS	marketing [129], psychology [130]
		Isomap	video processing [131], face recognition [132], wireless network [133]
		LLE	gesture recognition [134], speech recognition [135], hyperspectral images [136]
		LE	face recognition [137], [138], hyperspectral images [139]
		MVU	process monitoring [140], transfer learning [141]
		AE	speech processing [142], document processing [143], gene expression [144]
	Supervised	$t$ -SNE	cytometry [145], camera relocation [146], breast cancer [147], reinforcement learning [148]
		FLDA	face recognition [149], [150], gender recognition [151], action recognition [152], [153], EEG [124], [154], EMG [155], prototype selection [156]
		Kernel FLDA	face recognition [127], [157], palmprint recognition [158], prototype selection [159]
		SPCA	speech recognition [160], meteorology [161], gesture recognition [162], prototype selection [163]
		ML	face identification [164], action recognition [165], person re-identification [166]

TABLE II: Performance of feature selection and extraction methods on MNIST dataset.

		Method	# features	Accuracy
Feature Selection	Filters	CC	290	50.90%
		MI	400	68.44%
		$\chi^2$ Statistics	400	67.46%
		FCBF	15	31.10%
	Wrappers	SFS	400	86.67%
		PSO	403	59.42%
		GA	396	61.80%
Feature Extraction	Unsupervised	PCA	5	60.80%
		Kernel PCA	5	9.2%
		MDS	5	61.66%
		Isomap	5	75.30%
		LLE	5	65.56%
		LE	5	77.04%
		AE	5	83.20%
		$t$ -SNE	3	89.62%
	Supervised	FLDA	5	76.04%
		Kernel FLDA	5	21.34%
		SPCA	5	55.68%
		ML	5	56.98%
–	–	Original data	784	53.50%

## V. ILLUSTRATION AND EXPERIMENTS

### A. Comparison of Feature Selection and Extraction Methods

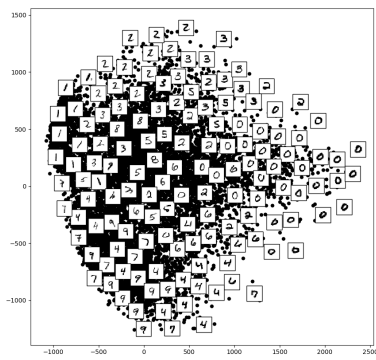
In order to compare the introduced methods, we tested them on a portion of MNIST dataset [178], i.e., the first 10000 and 5000 samples of training and testing sets, respectively. This dataset includes images of handwritten digits with size  $28 \times 28$ . Gaussian Naïve Bayes, as a simple classifier, is used for experiments in order to magnify the effectiveness of the feature selection and extraction methods.

The number of features in feature extraction is set to five (except  $t$ -SNE which we use three features for the sake of visualization). In some of feature selection methods, the number of selected features can be determined and we set it to 400 but some methods find out the best number of features themselves or based on a defined threshold. As reported in Table II, the accuracy of original data without applying any feature selection or extraction method is 53.50%. Except for CC, FCBF, Kernel PCA, and Kernel FLDA, all other methods have improved the performance. The  $t$ -SNE (state-of-the-art for visualization), AE with layers 784-50-50-50-50-784 (deep learning), and SFS have very good results. Non-linear methods such as Isomap, LLE, and LE have better results than linear methods (PCA and MDS) as expected. Kernel PCA, as was mentioned before, does not perform well in practice because of the choice of kernel (we used RBF kernel for it).

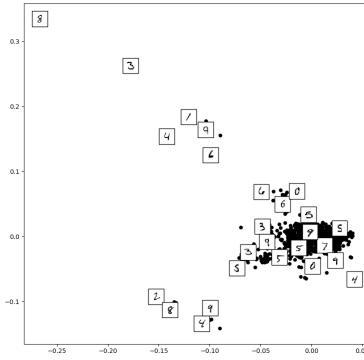
### B. Illustration of Embedded Space

For the sake of visualization, we applied feature extraction methods on the test set of MNIST dataset [178] (10,000 samples) and the samples are depicted in 2D embedded space in Fig. 1. As can be seen, the similar digits almost fall in the same clusters and different digits are separated acceptably. The AE, as a deep learning method, and the  $t$ -SNE, as the state-of-the-art for illustration show the best embedding among other methods. Kernel PCA and kernel FLDA do not have satisfactory results because of choice of kernels which are not optimum. The other methods have acceptable performance in embedding.

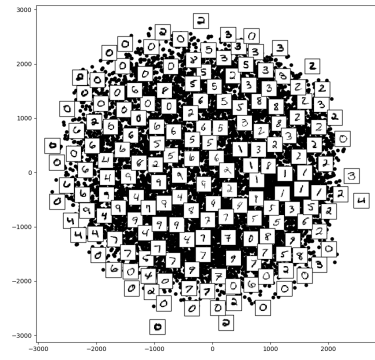
The performances of PCA and MDS are not very promising for this dataset in discriminating the digits because they



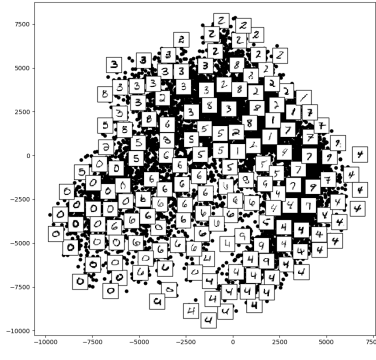
(a) PCA



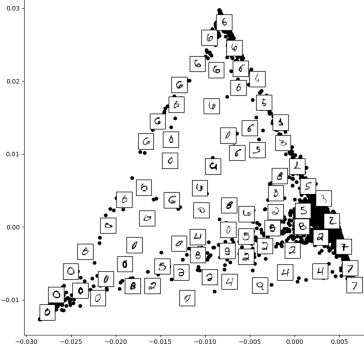
(b) Kernel PCA (RBF Kernel)



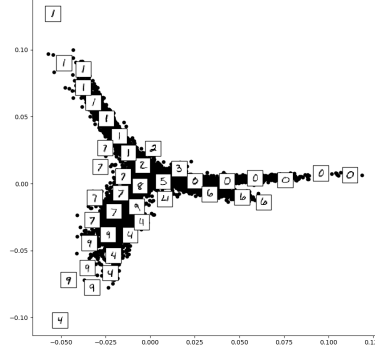
(c) MDS



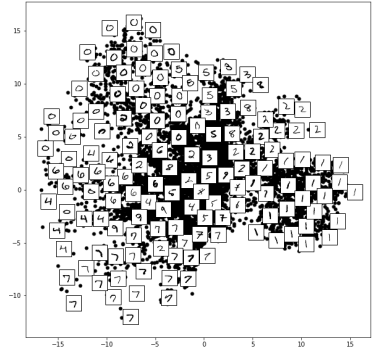
(d) Isomap



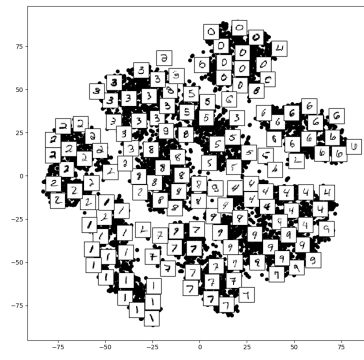
(e) LLE



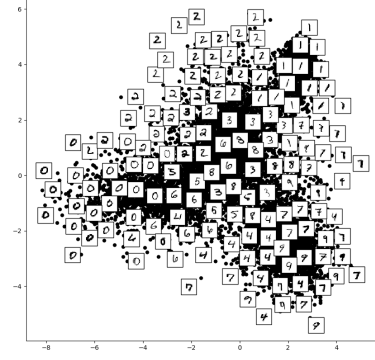
(f) LE



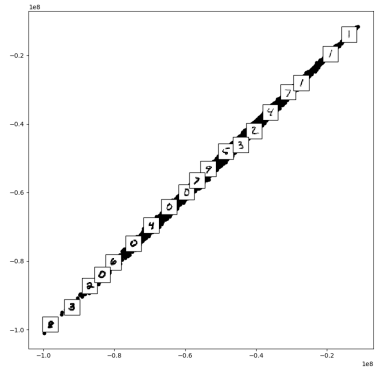
(g) AE



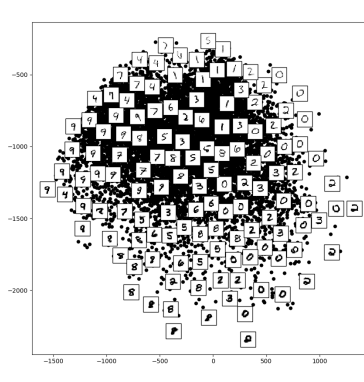
(h)  $t$ -SNE (After PCA)



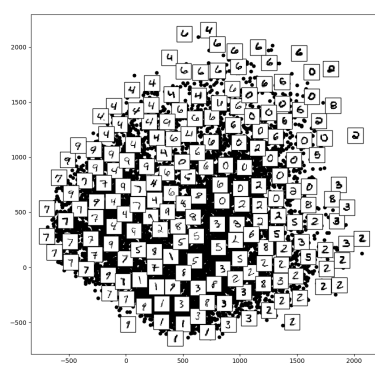
(i) FLDA



(j) Kernel FLDA (Linear Kernel)



(k) SPCA



(l) ML

Fig. 1: The MNIST dataset in embedded space obtained by different feature extraction methods.

are linear methods but the data sub-manifold is non-linear. Empirically, we have seen that the embedding of Isomap usually has several legs as in octopus. Two of octopus legs can be seen in Fig. 1, while for other datasets we might have more number of legs. The result of LLE is almost symmetric (symmetric triangle or square or etc) because in optimization of LLE which uses Eq. (22), the constraint is unit covariance [70]. Again empirically, we have seen that the embedding of LE usually includes some narrow string-like arms as also seen in Fig. 1. FLDA and SPCA have performed well because they make use of the class labels. ML has also performed well enough because it learns the optimum distance metric.

## VI. CONCLUSION

This paper discussed the motivations, theories, and differences of feature selection and extraction as a pre-processing stage for feature reduction. Some examples of the applications of the reviewed methods were also mentioned to show their usage in literature. Finally, the methods were tested on the MNIST dataset for comparison of their performances. Moreover, the embedded samples of MNIST dataset were illustrated for better interpretation.

## REFERENCES

- [1] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 Science and Information Conference*. IEEE, 2014, pp. 372–378.
- [2] C. O. S. Sorzano, J. Vargas, and A. P. Montano, "A survey of dimensionality reduction techniques," *arXiv preprint arXiv:1403.2877*, 2014.
- [3] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.
- [4] J. Miao and L. Niu, "A survey on feature selection," *Procedia Computer Science*, vol. 91, pp. 919–926, 2016.
- [5] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70–79, 2018.
- [6] Sudeshna Sarkar, IIT Kharagpur, "Introduction to machine learning course-feature selection." [Online]. Available: <https://www.youtube.com/watch?v=KTzXVnRlnw4&t=84s>
- [7] E. A. Guyon I, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, no. 3, pp. 1157–1182, 2003.
- [8] M. A. Hall, "Correlation-based feature selection for machine learning," *PhD thesis, University of Waikato, Hamilton*, 1999.
- [9] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," *IEEE Transactions on neural networks*, vol. 5, no. 4, pp. 537–550, 1994.
- [10] T. J. Cover TM, *Elements of Information Theory*. 2nd edn, Wiley-Interscience, New Jersey, 2006, vol. 2.
- [11] J. G. Kohavi R, "Wrappers for feature subset selection," *Elsevier*, no. 97, pp. 273–324, 1997.
- [12] T. Huijskens, "Mutual information-based feature selection," accessed: 2018-07-01. [Online]. Available: <https://thuijskens.github.io/2017/10/07/feature-selection/>
- [13] P. A. E. Jorge R. Vergara, "A review of feature selection methods based on mutual information," *Neural Comput and Applic*, pp. 175–186, 2014.
- [14] N. Nicolosiz, "Feature selection methods for text classification," Department of Computer Science, Rochester Institute of Technology, Tech. Rep., 2008.
- [15] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of Machine Learning Research*, vol. 3, pp. 1289–1305, 2003.
- [16] S. Fu and M. C. Desmarais, "Markov blanket based feature selection: A review of past decade," in *Proceedings of the World Congress on Engineering 2010 Vol I*. WCE, 2010.
- [17] I. Tsamardinos, C. F. Aliferis, A. R. Statnikov, and E. Statnikov, "Algorithms for large scale markov blanket discovery," in *FLAIRS conference*, vol. 2, 2003, pp. 376–380.
- [18] D. Margaritis and S. Thrun, "Bayesian network induction via local neighborhoods," in *Advances in neural information processing systems*, 2000, pp. 505–511.
- [19] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.
- [20] I. Tsamardinos, C. F. Aliferis, and A. Statnikov, "Time and sample efficient discovery of markov blankets and direct causal relations," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2003, pp. 673–678.
- [21] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial intelligence*, vol. 151, no. 1-2, pp. 155–176, 2003.
- [22] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.
- [23] Z. Zhao and H. Liu, "Searching for interacting features," in *international joint conference on artificial intelligence (IJCAI)*, vol. 7, 2007, pp. 1156–1161.
- [24] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.
- [25] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [26] M. B. Shirzad and M. R. Keyvanpour, "A feature selection method based on minimum redundancy maximum relevance for learning to rank," in *The 5th Conference on Artificial Intelligence and Robotics, IranOpen*, 2015, pp. 82–88.
- [27] D. W. Aha and R. L. Bankert, "A comparative evaluation of sequential feature selection algorithms," in *Learning from data*. Springer, 1996, pp. 199–206.
- [28] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern recognition letters*, vol. 15, no. 11, pp. 1119–1125, 1994.
- [29] I. J. A. H. S. M. Majdi M. Mafarja, Derar Eleyan, "Binary dragonfly algorithm for feature selection," *New Trends in Computing Sciences (ICTCS), 2017 International Conference*, pp. 12–17, 2017.
- [30] S. M. Majdi Mafarja, "Whale optimization approaches for wrapper feature selection," *Applied Soft Computing*, vol. 62, pp. 441–453, 2017.
- [31] J. K. R. Eberhart, "Particle swarm optimization," *Proc. of the IEEE International Conference on Neural Networks*, vol. 4, pp. 1942–1948, 1995.
- [32] A. Moraglio, C. Di Chio, and R. Poli, "Geometric particle swarm optimisation," in *European conference on genetic programming*. Springer, 2007, pp. 125–136.
- [33] E.-G. Talbi, L. Jourdan, J. Garcia-Nieto, and E. Alba, "Comparison of population based metaheuristics for feature selection: Application to microarray data classification," in *Computer Systems and Applications, 2008. AICCSA 2008. IEEE/ACS International Conference on*. IEEE, 2008, pp. 45–52.
- [34] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Systems, Man, and Cybernetics, 1997. Computational Cybernetics and Simulation., 1997 IEEE International Conference on*, vol. 5. IEEE, 1997, pp. 4104–4108.
- [35] S. J. Eiben, A.E., *Introduction to Evolutionary Computing, 2nd edn*. Springer Publishing Company, Incorporated (2015), 2015.
- [36] S. A.-N. A. R. S. M. R. Kazemi, M. M. S. Hoseini, "An evolutionary-based adaptive neurofuzzy inference system for intelligent shortterm load forecasting," *International Transactions in Operational Research*, vol. 21, pp. 311–326, 2014.
- [37] J. F.-C. E. S.-O. F. M.-d.-P. R. Urraca, A. Sanz-Gracia, "Improving hotel room demand forecasting with a hybrid ga-svr methodology based on skewed data transformation, feature selection and parsimony tuning," *Hybrid artificial intelligent systems*, vol. 9121, pp. 632–643, 2015.
- [38] L. J. Eshelman, "The chc adaptive search algorithm: how to have safe search when engaging in nontraditional genetic recombination," *Foundations of Genetic Algorithms*, vol. 1, pp. 265–283, 1991.
- [39] Y. Sun, C. Babbs, and E. Delp, "A comparison of feature selection methods for the detection of breast cancers in mammograms: Adap-

- tive sequential floating search vs. genetic algorithm,” *..*, vol. 6, pp. 6536–6539, 01 2005.
- [40] S. Jiang, K.-S. Chin, L. Wang, G. Qu, and K. L. Tsui, “Modified genetic algorithm-based feature selection combined with pre-trained deep neural network for demand forecasting in outpatient department,” *Expert Systems with Applications*, vol. 82, pp. 216–230, 2017.
  - [41] M. A. Carreira-Perpinán, “A review of dimension reduction techniques,” *Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09*, vol. 9, pp. 1–69, 1997.
  - [42] C. Fefferman, S. Mitter, and H. Narayanan, “Testing the manifold hypothesis,” *Journal of the American Mathematical Society*, vol. 29, no. 4, pp. 983–1049, 2016.
  - [43] C. J. Burges *et al.*, “Dimension reduction: A guided tour,” *Foundations and Trends® in Machine Learning*, vol. 2, no. 4, pp. 275–365, 2010.
  - [44] D. Engel, L. Hüttenberger, and B. Hamann, “A survey of dimension reduction methods for high-dimensional data analysis and visualization,” in *OASIS-OpenAccess Series in Informatics*, vol. 27. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2012.
  - [45] S. Liu, D. Maljovec, B. Wang, P.-T. Bremer, and V. Pascucci, “Visualizing high-dimensional data: Advances in the past decade,” *IEEE transactions on visualization and computer graphics*, vol. 23, no. 3, pp. 1249–1268, 2017.
  - [46] R. G. Baraniuk, V. Cevher, and M. B. Wakin, “Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 959–971, 2010.
  - [47] L. Cayton, “Algorithms for manifold learning,” *Univ. of California at San Diego Tech. Rep.*, pp. 1–17, 2005.
  - [48] Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.
  - [49] H. Strange and R. Zwiggelaar, *Open Problems in Spectral Dimensionality Reduction*. Springer, 2014.
  - [50] C. Robert, *Machine learning, a probabilistic perspective*. Taylor & Francis, 2014.
  - [51] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.
  - [52] S. Wold, K. Esbensen, and P. Geladi, “Principal component analysis,” *Chemometrics and intelligent laboratory systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
  - [53] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.
  - [54] K. Pearson, “Liii. on lines and planes of closest fit to systems of points in space,” *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.
  - [55] A. Ghodsi, “Dimensionality reduction: a short tutorial,” *Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, Canada. Tech. Rep.*, pp. 1–25, 2006.
  - [56] A. Ghodsi, “Data visualization course, winter 2017,” accessed: 2018-01-01. [Online]. Available: <https://www.youtube.com/watch?v=L-pQtGm3VS8&list=PLLeuLRPyt1HzQoXEhtNuYTmd0aNQvtyAK>
  - [57] B. Ghogh, F. Karray, and M. Crowley, “Eigenvalue and generalized eigenvalue problems: Tutorial,” *arXiv preprint arXiv:1903.11240*, 2019.
  - [58] T. P. Minka, “Automatic choice of dimensionality for pca,” in *Advances in neural information processing systems*, 2001, pp. 598–604.
  - [59] R. B. Cattell, “The scree test for the number of factors,” *Multivariate behavioral research*, vol. 1, no. 2, pp. 245–276, 1966.
  - [60] B. Schölkopf, A. Smola, and K.-R. Müller, “Kernel principal component analysis,” in *International Conference on Artificial Neural Networks*. Springer, 1997, pp. 583–588.
  - [61] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.
  - [62] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” *The annals of statistics*, pp. 1171–1220, 2008.
  - [63] D. L. Donoho, “High-dimensional data analysis: The curses and blessings of dimensionality,” *AMS math challenges lecture*, vol. 1, no. 2000, pp. 1–33, 2000.
  - [64] B. Schölkopf, S. Mika, A. Smola, G. Rätsch, and K.-R. Müller, “Kernel PCA pattern reconstruction via approximate pre-images,” in *ICANN 98*. Springer, 1998, pp. 147–152.
  - [65] M. A. Cox and T. F. Cox, “Multidimensional scaling,” in *Handbook of data visualization*. Springer, 2008, pp. 315–347.
  - [66] J. C. Gower, “Some distance properties of latent root and vector methods used in multivariate analysis,” *Biometrika*, vol. 53, no. 3-4, pp. 325–338, 1966.
  - [67] J. B. Tenenbaum, V. De Silva, and J. C. Langford, “A global geometric framework for nonlinear dimensionality reduction,” *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.
  - [68] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to algorithms*. MIT press, 2009.
  - [69] L. C. Marsh and D. R. Cormier, *Spline regression models*. Sage, 2001, vol. 137.
  - [70] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.
  - [71] L. K. Saul and S. T. Roweis, “Think globally, fit locally: unsupervised learning of low dimensional manifolds,” *Journal of machine learning research*, vol. 4, no. Jun, pp. 119–155, 2003.
  - [72] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural computation*, vol. 15, no. 6, pp. 1373–1396, 2003.
  - [73] J. H. Ham, D. D. Lee, S. Mika, and B. Schölkopf, “A kernel view of the dimensionality reduction of manifolds,” in *International Conference on Machine Learning*, 2004.
  - [74] H. Strange and R. Zwiggelaar, “Spectral dimensionality reduction,” in *Open Problems in Spectral Dimensionality Reduction*. Springer, 2014, pp. 7–22.
  - [75] K. Q. Weinberger, F. Sha, and L. K. Saul, “Learning a kernel matrix for nonlinear dimensionality reduction,” in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 106.
  - [76] K. Q. Weinberger and L. K. Saul, “Unsupervised learning of image manifolds by semidefinite programming,” *International journal of computer vision*, vol. 70, no. 1, pp. 77–90, 2006.
  - [77] —, “An introduction to nonlinear dimensionality reduction by maximum variance unfolding,” in *AAAI*, vol. 6, 2006, pp. 1683–1686.
  - [78] L. Vandenberghe and S. Boyd, “Semidefinite programming,” *SIAM review*, vol. 38, no. 1, pp. 49–95, 1996.
  - [79] L. van der Maaten, E. Postma, and J. van den Herik, “Dimensionality reduction: A comparative review,” *Tilburg centre for Creative Computing. Tilburg University. Tech. Rep.*, pp. 1–36, 2009.
  - [80] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
  - [81] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
  - [82] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning internal representations by error propagation,” *California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep.*, 1985.
  - [83] G. E. Hinton, “Deep belief networks,” *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.
  - [84] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.
  - [85] S. Ioffe and C. Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in *Proceedings of the 32nd international conference on machine learning (ICML)*, 2015, pp. 448–456.
  - [86] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
  - [87] L. v. d. Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.
  - [88] G. E. Hinton and S. T. Roweis, “Stochastic neighbor embedding,” in *Advances in neural information processing systems*, 2003, pp. 857–864.
  - [89] W. S. Gosset (Student), “The probable error of a mean,” *Biometrika*, pp. 1–25, 1908.
  - [90] S. Kullback, *Information theory and statistics*. Courier Corporation, 1997.
  - [91] L. van der Maaten, “Learning a parametric embedding by preserving local structure,” in *Artificial Intelligence and Statistics*, 2009, pp. 384–391.
  - [92] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” *Annals of eugenics*, vol. 7, no. 2, pp. 179–188, 1936.

- [93] M. Welling, "Fisher linear discriminant analysis," University of Toronto, Toronto, Ontario, Canada, Tech. Rep., 2005.
- [94] Y. Xu and G. Lu, "Analysis on Fisher discriminant criterion and linear separability of feature space," in *2006 International Conference on Computational Intelligence and Security*, vol. 2. IEEE, 2006, pp. 1671–1676.
- [95] K. P. Murphy, *Machine learning: a probabilistic perspective*. MIT press, 2012.
- [96] J. Yang and J.-y. Yang, "Why can LDA be performed in PCA transformed space?" *Pattern recognition*, vol. 36, no. 2, pp. 563–566, 2003.
- [97] B. Ghoghj and H. Mohammadzade, "Automatic extraction of key-poses and key-joints for action recognition using 3d skeleton data," in *2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP)*. IEEE, 2017, pp. 164–170.
- [98] H. F. Eid, A. E. Hassanien, T.-h. Kim, and S. Banerjee, "Linear correlation-based feature selection for network intrusion detection model," in *Advances in Security of Information and Communication Networks*. Springer, 2013, pp. 240–248.
- [99] M. Ciesielczyk, "Using mutual information for feature selection in programmatic advertising," in *INnovations in Intelligent SysTems and Applications (INISTA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 290–295.
- [100] B. Fish, A. Khan, N. H. Chehade, C. Chien, and G. Pottie, "Feature selection based on mutual information for human activity recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1729–1732.
- [101] R. Abraham, J. B. Simha, and S. Iyengar, "Medical datamining with a new algorithm for feature selection and naïve bayesian classifier," in *Information Technology (ICIT 2007). 10th International Conference on*. IEEE, 2007, pp. 44–49.
- [102] A. Moh'd A Mesleh, "Chi square feature extraction based svms arabic language text categorization system," *Journal of Computer Science*, vol. 3, no. 6, pp. 430–435, 2007.
- [103] T. Basu and C. Murthy, "Effective text classification by a supervised feature selection approach," in *2012 IEEE 12th International Conference on Data Mining Workshops*. IEEE, 2012, pp. 918–925.
- [104] H. Zeng and Y.-M. Cheung, "A new feature selection method for gaussian mixture clustering," *Pattern Recognition*, vol. 42, no. 2, pp. 243–250, 2009.
- [105] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical ip traffic flow classification," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 5, pp. 5–16, 2006.
- [106] Y. Liu and M. Schumann, "Data mining feature selection for credit scoring models," *Journal of the Operational Research Society*, vol. 56, no. 9, pp. 1099–1108, 2005.
- [107] D. Rodriguez, R. Ruiz, J. Cuadrado-Gallego, J. Aguilar-Ruiz, and M. Garre, "Attribute selection in software engineering datasets for detecting fault modules," in *Software Engineering and Advanced Applications, 2007. 33rd EUROMICRO Conference on*. IEEE, 2007, pp. 418–423.
- [108] S. H. Huang, L. R. Wulsin, H. Li, and J. Guo, "Dimensionality reduction for knowledge discovery in medical claims database: application to antidepressant medication utilization study," *Computer methods and programs in biomedicine*, vol. 93, no. 2, pp. 115–123, 2009.
- [109] S. Liu, X. Chen, W. Liu, J. Chen, Q. Gu, and D. Chen, "Fecar: A feature selection framework for software defect prediction," in *Computer Software and Applications Conference (COMPSAC), 2014 IEEE 38th Annual*. IEEE, 2014, pp. 426–435.
- [110] A. W. Moore and D. Zuev, "Internet traffic classification using bayesian analysis techniques," in *ACM SIGMETRICS Performance Evaluation Review*, vol. 33, no. 1. ACM, 2005, pp. 50–60.
- [111] R. S. Baker, "Modeling and understanding students' off-task behavior in intelligent tutoring systems," in *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 2007, pp. 1059–1068.
- [112] S. Kandula, R. Mahajan, P. Verkaik, S. Agarwal, J. Padhye, and P. Bahl, "Detailed diagnosis in enterprise networks," *ACM SIGCOMM Computer Communication Review*, vol. 39, no. 4, pp. 243–254, 2009.
- [113] L. Koc, T. A. Mazzuchi, and S. Sarkani, "A network intrusion detection system based on a hidden naïve bayes multiclass classifier," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13 492–13 500, 2012.
- [114] G. Wang, Q. Song, H. Sun, X. Zhang, B. Xu, and Y. Zhou, "A feature subset selection algorithm automatic recommendation method," *Journal of Artificial Intelligence Research*, vol. 47, pp. 1–34, 2013.
- [115] B. Remeseiro, V. Bolon-Canedo, D. Peteiro-Barral, A. Alonso-Betanzos, B. Guijarro-Berdinas, A. Mosquera, M. G. Penedo, and N. Sánchez-Marono, "A methodology for improving tear film lipid layer classification," *IEEE journal of biomedical and health informatics*, vol. 18, no. 4, pp. 1485–1493, 2014.
- [116] X. Jin, E. W. Ma, L. L. Cheng, and M. Pecht, "Health monitoring of cooling fans based on mahalanobis distance with mrmr feature selection," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 8, pp. 2222–2229, 2012.
- [117] A. Idris, A. Khan, and Y. S. Lee, "Intelligent churn prediction in telecom: employing mrmr feature selection and rotboost based ensemble classification," *Applied intelligence*, vol. 39, no. 3, pp. 659–672, 2013.
- [118] M. Radovic, M. Ghalwash, N. Filipovic, and Z. Obradovic, "Minimum redundancy maximum relevance feature selection approach for temporal gene expression data," *BMC bioinformatics*, vol. 18, no. 1, p. 9, 2017.
- [119] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 2, pp. 153–158, 1997.
- [120] T. Rückstieß, C. Osendorfer, and P. van der Smagt, "Sequential feature selection for classification," in *Australasian Joint Conference on Artificial Intelligence*. Springer, 2011, pp. 132–141.
- [121] E. Alba, J. Garcia-Nieto, L. Jourdan, and E.-G. Talbi, "Gene selection in cancer classification using pso/svm and ga/svm hybrid algorithms," in *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on*. IEEE, 2007, pp. 284–290.
- [122] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," in *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*. IEEE, 1991, pp. 586–591.
- [123] M. Ahmad and S.-W. Lee, "Hmm-based human action recognition using multiview image sequences," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, 2006, pp. 263–266.
- [124] A. Subasi and M. I. Gursoy, "Eeg signal classification using pca, ica, lda and support vector machines," *Expert systems with applications*, vol. 37, no. 12, pp. 8659–8666, 2010.
- [125] H. Mohammadzade, B. Ghoghj, S. Faezi, and M. Shabany, "Critical object recognition in millimeter-wave images with robustness to rotation and scale," *JOSA A*, vol. 34, no. 6, pp. 846–855, 2017.
- [126] K. I. Kim, K. Jung, and H. J. Kim, "Face recognition using kernel principal component analysis," *IEEE signal processing letters*, vol. 9, no. 2, pp. 40–42, 2002.
- [127] M.-H. Yang, "Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods," in *fgv*. IEEE, 2002, p. 0215.
- [128] S. W. Choi, C. Lee, J.-M. Lee, J. H. Park, and I.-B. Lee, "Fault detection and identification of nonlinear processes based on kernel pca," *Chemometrics and intelligent laboratory systems*, vol. 75, no. 1, pp. 55–67, 2005.
- [129] L. G. Cooper, "A review of multidimensional scaling in marketing research," *Applied Psychological Measurement*, vol. 7, no. 4, pp. 427–450, 1983.
- [130] S. L. Robinson and R. J. Bennett, "A typology of deviant workplace behaviors: A multidimensional scaling study," *Academy of management journal*, vol. 38, no. 2, pp. 555–572, 1995.
- [131] R. Pless, "Image spaces and video trajectories: Using isomap to explore video sequences," in *ICCV*, vol. 3, 2003, pp. 1433–1440.
- [132] M.-H. Yang, "Face recognition using extended isomap," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 2. IEEE, 2002, pp. II–II.
- [133] C. Wang, J. Chen, Y. Sun, and X. Shen, "Wireless sensor networks localization with isomap," in *Communications, 2009. ICC'09. IEEE International Conference on*. IEEE, 2009, pp. 1–5.
- [134] S. S. Ge, Y. Yang, and T. H. Lee, "Hand gesture recognition and tracking based on distributed locally linear embedding," *Image and Vision Computing*, vol. 26, no. 12, pp. 1607–1620, 2008.
- [135] V. Jain and L. K. Saul, "Exploratory analysis and visualization of speech and music by locally linear embedding," in *Acoustics, Speech,*

- and Signal Processing, 2004. *Proceedings (ICASSP'04)*. IEEE International Conference on, vol. 3. IEEE, 2004, pp. iii–984.
- [136] D. Kim and L. Finkel, “Hyperspectral image processing using locally linear embedding,” in *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference on*. IEEE, 2003, pp. 316–319.
- [137] W. Luo, “Face recognition based on laplacian eigenmaps,” in *Computer Science and Service System (CSSS), 2011 International Conference on*. IEEE, 2011, pp. 416–419.
- [138] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, “Face recognition using laplacianfaces,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 328–340, 2005.
- [139] B. Hou, X. Zhang, Q. Ye, and Y. Zheng, “A novel method for hyperspectral image classification based on laplacian eigenmap pixels distribution-flow,” *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 6, no. 3, pp. 1602–1618, 2013.
- [140] J.-D. Shao and G. Rong, “Nonlinear process monitoring based on maximum variance unfolding projections,” *Expert Systems with Applications*, vol. 36, no. 8, pp. 11 332–11 340, 2009.
- [141] S. J. Pan, J. T. Kwok, and Q. Yang, “Transfer learning via dimensionality reduction,” in *AAAI*, vol. 8, 2008, pp. 677–682.
- [142] L. Deng, M. L. Seltzer, D. Yu, A. Acero, A.-r. Mohamed, and G. Hinton, “Binary coding of speech spectrograms using a deep auto-encoder,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
- [143] J. Li, T. Luong, and D. Jurafsky, “A hierarchical neural autoencoder for paragraphs and documents,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, vol. 1, 2015, pp. 1106–1115.
- [144] D. Chicco, P. Sadowski, and P. Baldi, “Deep autoencoder neural networks for gene ontology annotation predictions,” in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2014, pp. 533–540.
- [145] C. Chester and H. T. Maecker, “Algorithmic tools for mining high-dimensional cytometry data,” *The Journal of Immunology*, vol. 195, no. 3, pp. 773–779, 2015.
- [146] A. Kendall, M. Grimes, and R. Cipolla, “Posenet: A convolutional network for real-time 6-dof camera relocation,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.
- [147] T. Araújo, G. Aresta, E. Castro, J. Rouco, P. Aguiar, C. Eloy, A. Polónia, and A. Campilho, “Classification of breast cancer histology images using convolutional neural networks,” *PloS one*, vol. 12, no. 6, p. e0177544, 2017.
- [148] T. Zahavy, N. Ben-Zrihem, and S. Mannor, “Graying the black box: Understanding dqns,” in *International Conference on Machine Learning*, 2016, pp. 1899–1908.
- [149] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. fisherfaces: recognition using class specific linear projection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.
- [150] H. Mohammadzade, A. Sayyafan, and B. Ghogh, “Pixel-level alignment of facial images for high accuracy recognition using ensemble of patches,” *JOSA A*, vol. 35, no. 7, pp. 1149–1159, 2018.
- [151] B. Ghogh, S. B. Shouraki, H. Mohammadzade, and E. Iranmehr, “A fusion-based gender recognition method using facial images,” in *Electrical Engineering (ICEE), Iranian Conference on*. IEEE, 2018, pp. 1493–1498.
- [152] B. Ghogh, H. Mohammadzade, and M. Mokari, “Fisherposes for human action recognition using kinect sensor data,” *IEEE Sensors Journal*, vol. 18, no. 4, pp. 1612–1627, 2017.
- [153] M. Mokari, H. Mohammadzade, and B. Ghogh, “Recognizing involuntary actions from 3d skeleton data using body states,” *Scientia Iranica*, 2018.
- [154] A. Malekmohammadi, H. Mohammadzade, A. Chamanzar, M. Shabany, and B. Ghogh, “An efficient hardware implementation for a motor imagery brain computer interface system,” *Scientia Iranica*, vol. 26, pp. 72–94, 2019.
- [155] A. Phinyomark, H. Hu, P. Phukpattaranont, and C. Limsakul, “Application of linear discriminant analysis in dimensionality reduction for hand motion classification,” *Measurement Science Review*, vol. 12, no. 3, pp. 82–89, 2012.
- [156] B. Ghogh and M. Crowley, “Principal sample analysis for data reduction,” in *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE, 2018, pp. 350–357.
- [157] M.-H. Yang, “Face recognition using kernel fisherfaces,” 2006, uS Patent 7,054,468.
- [158] Y. Wang and Q. Ruan, “Kernel fisher discriminant analysis for palmprint recognition,” in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 4. IEEE, 2006, pp. 457–460.
- [159] B. Ghogh, “Principal sample analysis for data ranking,” in *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019*. Springer, 2019.
- [160] P. Fewzee and F. Karray, “Dimensionality reduction for emotional speech recognition,” in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*. IEEE, 2012, pp. 532–537.
- [161] A. Sarhadi, D. H. Burn, G. Yang, and A. Ghodsi, “Advances in projection of climate change impacts using supervised nonlinear dimensionality reduction techniques,” *Climate dynamics*, vol. 48, no. 3-4, pp. 1329–1351, 2017.
- [162] A.-A. Samadani, A. Ghodsi, and D. Kulić, “Discriminative functional analysis of human movements,” *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1829–1839, 2013.
- [163] B. Ghogh and M. Crowley, “Instance ranking and numerosity reduction using matrix decomposition and subspace learning,” in *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019*. Springer, 2019.
- [164] M. Guillaumin, J. Verbeek, and C. Schmid, “Is that you? metric learning approaches for face identification,” in *ICCV 2009-International Conference on Computer Vision*. IEEE, 2009, pp. 498–505.
- [165] D. Tran and A. Sorokin, “Human activity recognition with metric learning,” in *European conference on computer vision*. Springer, 2008, pp. 548–561.
- [166] D. Yi, Z. Lei, S. Liao, and S. Z. Li, “Deep metric learning for person re-identification,” in *Pattern Recognition (ICPR), 2014 22nd International Conference on*. IEEE, 2014, pp. 34–39.
- [167] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, “Fisher discriminant analysis with kernels,” in *Neural networks for signal processing IX, 1999. Proceedings of the 1999 IEEE signal processing society workshop*. IEEE, 1999, pp. 41–48.
- [168] E. Bair and R. Tibshirani, “Semi-supervised methods to predict patient survival from gene expression data,” *PLoS biology*, vol. 2, no. 4, p. e108, 2004.
- [169] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, “Prediction by supervised principal components,” *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 119–137, 2006.
- [170] P. Daniušis, P. Vaitkus, and L. Petkevičius, “Hilbert–Schmidt component analysis,” *Proc. of the Lithuanian Mathematical Society, Ser. A*, vol. 57, pp. 7–11, 2016.
- [171] E. Barshan, A. Ghodsi, Z. Azimifar, and M. Z. Jahromi, “Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds,” *Pattern Recognition*, vol. 44, no. 7, pp. 1357–1371, 2011.
- [172] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, “Measuring statistical dependence with Hilbert-Schmidt norms,” in *International conference on algorithmic learning theory*. Springer, 2005, pp. 63–77.
- [173] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, “Distance metric learning with application to clustering with side-information,” in *Advances in neural information processing systems*, 2003, pp. 521–528.
- [174] J. Peltonen, A. Klami, and S. Kaski, “Improved learning of Riemannian metrics for exploratory analysis,” *Neural Networks*, vol. 17, no. 8-9, pp. 1087–1100, 2004.
- [175] B. Alipanahi, M. Biggs, and A. Ghodsi, “Distance metric learning vs. fisher discriminant analysis,” in *Proceedings of the 23rd national conference on Artificial intelligence*, vol. 2, 2008, pp. 598–603.
- [176] A. Globerson and S. T. Roweis, “Metric learning by collapsing classes,” in *Advances in neural information processing systems*, 2006, pp. 451–458.
- [177] A. Ghodsi, D. F. Wilkinson, and F. Southey, “Improving embeddings by flexible exploitation of side information,” in *IJCAI*, 2007, pp. 810–816.
- [178] Y. LeCun, C. Cortes, and C. J. Burges, “the mnist database of handwritten digits,” accessed: 2018-07-01. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>