# Fractal Autoencoders for Feature Selection

**Xinxing Wu, Qiang Cheng***
University of Kentucky, Lexington, Kentucky, USA
xinxingwu@gmail.com, qiang.cheng@uky.edu

## Abstract

Feature selection reduces the dimensionality of data by identifying a subset of the most informative features. In this paper, we propose an innovative framework for unsupervised feature selection, called fractal autoencoders (FAE). It trains a neural network to pinpoint informative features for global exploring of representability and for local excavating of diversity. Architecturally, FAE extends autoencoders by adding a one-to-one scoring layer and a small sub-neural network for feature selection in an unsupervised fashion. With such a concise architecture, FAE achieves state-of-the-art performances; extensive experimental results on fourteen datasets, including very high-dimensional data, have demonstrated the superiority of FAE over existing contemporary methods for unsupervised feature selection. In particular, FAE exhibits substantial advantages on gene expression data exploration, reducing measurement cost by about 15% over the widely used L1000 landmark genes. Further, we show that the FAE framework is easily extensible with an application.

## Introduction

High-dimensional data is pervasive in almost every area of modern data science (Clarke et al. 2008; Blum, Hopcroft, and Kannan 2020). Dealing with high-dimensional data is challenging due to the known phenomenon – the curse of dimensionality (Bellman 1957). In numerous applications, principal component analysis (PCA) (Pearson 1901) and autoencoders (AE) (Rumelhart, Hinton, and Williams 1985; Ballard 1987), two traditional and simple approaches, are typically used for dimensionality reduction. For example, PCA is adopted to reduce the dimensions of gene expression data before the extraction of the samples' rhythmic structures (Anafi et al. 2017); AE are employed to process high-dimensional datasets prior to clustering (Xie, Girshick, and Farhadi 2016) and subspace learning (Ji et al. 2017). Despite their widespread usage, the interpretation of lower-dimensional feature spaces produced by PCA and AE is not straightforward, because these feature spaces are different from the original feature space. In contrast to PCA and AE, feature selection allows for ready interpretability with the input features, by identifying and retaining a subset of important features directly from the original feature space (Guyon and Elisseeff 2003).

There exist various feature selection approaches. According to whether labels are used, they can be categorized as supervised (Cheng, Zhou, and Cheng 2011), semi-supervised, and unsupervised methods (Alelyani, Tang, and Liu 2013). Unsupervised approaches have potentially extensive applications, since they do not require labels that can be rare or expensive to obtain. A variety of techniques for unsupervised feature selection have been proposed, e.g., Laplacian score (LS) (He, Cai, and Niyogi 2005) and concrete autoencoders (CAE) (Abid, Balin, and Zou 2019). While often used, the existing approaches may still exhibit suboptimal performance in downstream learning tasks on many datasets, which can be seen, e.g., in Table 3. There are two major reasons that cause such under-performance. First, the space to search for potentially important subsets of features in the absence of the guidance by labels is often very large, which renders unsupervised feature selection to be like finding a needle in a haystack. Second, it is necessary, yet challenging, to take account of the inter-feature interactions. Ideally, the selected features should be globally representative and as diverse as possible. If the selected features are all important yet highly correlated, they may be capable of representing only partial data, and thus they would hardly comprise a globally representative feature subset. For example, if a pixel of a natural image is important, then some neighboring ones are also likely to be so because of the typical spatial dependence in images; thus, to select diverse, salient features to represent the overall contents, if a pixel is important and selected, those neighboring pixels of high correlations with it should not be included into the feature subset. Existing unsupervised approaches have limited abilities to simultaneously explore the large search space for features that can represent the overall contents and take into account the diversity, which is reflected by the inter-correlation of features, thus leading to suboptimal performances.

To overcome these difficulties, in this paper we propose a novel unsupervised feature selection framework, called fractal autoencoders (FAE). It trains a neural network (NN) to identify potentially informative features for representing the contents globally; simultaneously, it exploits a dependence sub-NN to select a subset locally from the globally informative features to examine their diversity, which is efficiently

*Corresponding author

Copyright © 2021, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

measured by their abilities to reconstruct the original data. In this way, the sub-NN enables FAE to effectively screen out the highly correlated features; the global AE component of FAE turns out to play a crucial role of regularization to stabilize the feature selecting process, aside from its standard role of feature extraction. With our new architecture, FAE merges feature selection and feature extraction into one model, facilitating the identification of a subset of the most representative and diverse input features. To illustrate the extensive ability of the FAE framework, we use it to derive an $h$-Hierarchy FAE ($h$-HFAE) application to identity multiple subsets of important features.

In summary, our main contributions include:

- We propose a novel framework, FAE, for feature selection to meet the challenges of existing methods. It combines global exploration of representative features and local excavation of their diversity, thereby enhancing the generalization of the selected feature subset and accounting for inter-feature correlations.
- As our framework can be readily applicable or extensible to other tasks, we show an application to identity multiple hierarchical subsets of salient features simultaneously.
- We validate FAE with extensive experiments on fourteen real datasets. Although simple, it demonstrates state-of-the-art performance for reconstruction on many benchmarking datasets. It also yields superior performance in a downstream learning task of classification on most of the benchmarking datasets. As a biological application, FAE reduces gene expression measurements by about 15% compared with L1000 landmark genes. Further, FAE exhibits more stable performance on varying numbers of selected features than contemporary methods.

The notations and definitions are given as follows. Let $n$, $m$, $k$, and $d$ be the numbers of samples, features, selected features, and reduced dimensions, respectively. Let $\mathbf{X} \in \mathbb{R}^{n\times m}$ be a matrix containing the input data. A bold capital letter such as $\mathbf{W}$ denotes a matrix; a lowercase bold capital letter such as $\mathbf{w}$ denotes a vector; Diag($\mathbf{w}$) represents a diagonal matrix with the diagonal $\mathbf{w}$; $\mathbf{w}^{\max_k}$ is an operation to keep the $k$ largest entries of $\mathbf{w}$ while making other entries 0. $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm.

The remaining of the paper is organized as follows. We first discuss the related work, then present our proposed approach, followed by extensive experiments. Finally, we apply FAE to identify multiple subsets of informative features.

## Related Work

A variety of feature selection approaches have been proposed. They are usually classified into four categories (Alelyani, Tang, and Liu 2013; Li et al. 2017): filter methods, which are independent of learning models; wrapper methods, which rely on learning models for selection criteria; embedder approaches, which embed the feature selection into learning models to also achieve model fitting simultaneously; hybrid approaches, which are a combination of more than one of above three. Alternatively, the approaches are categorized as supervised, semi-supervised, and unsupervised methods according to whether label information is utilized. Unsupervised feature selection has potentially broad applications because it requires no label information (Peng et al. 2016, 2017); yet, it is also arguably more challenging due to the lack of labels to guide the identification of relevant features. In this paper, we focus on unsupervised feature selection and briefly review typical methods below.

LS (He, Cai, and Niyogi 2005) is a filter method that uses the nearest neighbor graph to model the local geometric structures of the data. By selecting the features which are locally the smoothest on the graph, LS focuses on the local property yet neglects the global structure. SPEC (Zhao and Liu 2007) is a filter method based on general similarity matrix. It employs the spectrum of the graph to measure feature relevance and unifies supervised and unsupervised feature selection. Principal feature analysis (PFA) (Lu et al. 2007) utilizes the structure of the principal components of a set of features to select the subset of relevant features. It can be regarded as a wrapper method to optimize the PC coefficients, and it mainly focuses on globality. Multi-cluster feature selection (MCFS) (Cai, Zhang, and He 2010) selects a subset of features to cover the multi-cluster structure of the data, where spectral analysis is used to find the inter-relationship between different features. Unsupervised discriminative feature selection (UDFS) (Yang et al. 2011) incorporates the discriminative analysis and $\ell_{2,1}$ regularization to identify the most useful features. Nonnegative discriminative feature selection (NDFS) (Li et al. 2012) jointly learns the cluster labels and feature selection matrix to select discriminative features. It uses a nonnegative constraint on the class indicator to learn cluster labels and adopts an $\ell_{2,1}$ constraint to reduce the redundant or noisy features. Infinite feature selection (Inf-FS) (Roffo, Melzi, and Cristani 2015) implements feature selection by taking into account all the possible feature subsets as paths on a graph, and it is also a filter method.

Recently, a few AE-based feature selection methods have been developed. Autoencoder feature selector (AEFS) (Han et al. 2018) combines autoencoders regression and $\ell_{2,1}$ regularization on the weights of the encoder to obtain a subset of useful features. It exploits both linear and nonlinear information in the features. Agnostic feature selection (AgnoS) (Doquet and Sebag 2019) adopts AE with explicit objective function regularizations, such as the $\ell_{2,1}$ norm on the weights of the first layer of AE (AgnoS-W), $\ell_{2,1}$ norm on the gradient of the encoder (AgnoS-G), and $\ell_1$ norm on the slack variables that constitute the first layer of AE (AgnoS-S), to implement feature selection. AgnoS-S is the best of the three, so in this study we will compare our approach with AgnoS-S. CAE (Abid, Balin, and Zou 2019) replaces the first hidden layer of AE with a “concrete selector” layer, which is the relaxation of a discrete distribution called concrete distribution (Maddison, Mnih, and Teh 2017), and then it picks the features with an extremely high probability of connecting to the nodes of the concrete selection layer. The parameters of this layer are estimated by the reparametrization trick (Kingma and Welling 2014). CAE reports superior performance over other competing methods.

MCFS, UDFS, NDFS, AEFS, AgnoS, and CAE can be all regarded as embedded approaches. Though our proposed FAE model also embeds the feature selection into AE, which looks similar to AgnoS, AEFS, and CAE, it essentially differs from these existing methods: AEFS and AgnoS mainly depend on exploiting sparsity norm regularizations such as $\ell_{2,1}$ and $\ell_1$ on the weights of AE to select features, which do not consider diversity; in contrast, FAE consists of two NNs, and it innovatively adopts a sub-NN to explicitly impose the desired diversity requirement on informative features. CAE adopts a probability distribution on the first layer of AE and selects features by their parameters. However, several neurons in the concrete selector layer may potentially select the same or redundant features, and the training requires that the average of the maximum probability of connecting to these neurons in the concrete selection layer exceed a prespecified threshold close to 1, which may be hard to attain for high-dimensional datasets; meanwhile, the second and third top features at different nodes of the concrete selector layer may be insignificant because of their trivial average probability. These potential drawbacks can limit the performance of CAE. In contrast, the proposed FAE does not depend on any probability distribution; rather, its sub-NN, with the guidance by the global-NN, directly pinpoints a subset of selected features, which makes FAE concise in architecture and easily applicable to different tasks.

## Proposed Approach

In this section, we will present the architecture and formulation of FAE.

### Overview of Our Approach

The architecture of our FAE approach is depicted in Figure 1. It enlists the AE architecture as a basic building block; yet, its structure is particularly tailored to feature selection. In the following, we will explain the architecture and its components in detail.

### Formalization of Autoencoders

For AE, we formalize it as follows:

$$\min_{f,g} \|\mathbf{X} - f(g(\mathbf{X}))\|_{\mathrm{F}}^2, \tag{1}$$

where $g$ is an encoder, and $f$ is a decoder. $g(\mathbf{X})$ embeds the input data into a latent space $\mathbb{R}^{n \times d}$, where $d$ is the dimension of the bottleneck layer of AE. Taking MNIST as an example, for $d = 49$, we visualize the encoded samples in Figure 2 (b). After being transformed, either nonlinearly or linearly, from the original space, the contents of each sample are not visually meaningful in the latent space.

### Formalization of Unsupervised Feature Selection

Feature selection is to identify a subset of informative features in the original feature space, and it can be formalized as follows:

$$\min_{S^k, H} \|H(\mathbf{X}_{S^k}) - \mathbf{X}\|_{\mathrm{F}}^2, \tag{2}$$

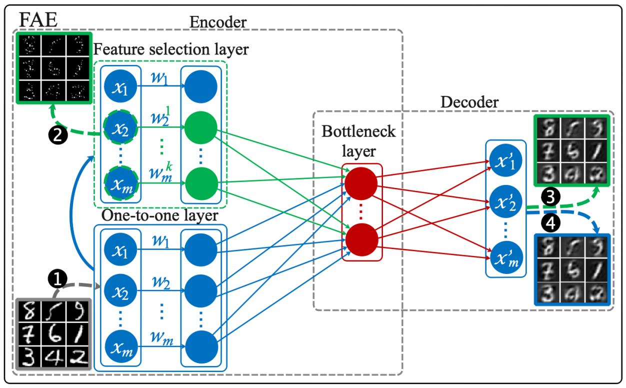


Figure 1: The architecture of FAE. During training, the global NN (with one-to-one layer) and its dependence sub-NN (with feature selection layer) are used to optimize (3); during testing, only the trained sub-NN is used to select features and reconstruct the data. Potentially, FAE implements feature extraction. The presented quantifies are: 1) input; 2) feature selection result; 3) reconstruction based on the selected features, i.e., $f(g(\mathbf{X}\mathbf{W}_{\mathrm{I}}^{\max_k}))$; 4) reconstruction from the one-to-one layer, i.e., $f(g(\mathbf{X}\mathbf{W}_{\mathrm{I}}))$.

where $S^k$ denotes the subset of the specified $k$ features , $\mathbf{X}_{S^k}$ is the derived data set from $\mathbf{X}$ based on $S^k$, and $H$ denotes a mapping from the space spanned by $\mathbf{X}_{S^k}$ to $\mathbb{R}^{n \times m}$ in the absence of information about labels. The optimization problem in (2) is NP-hard (Natarajan 1995; Hamo and Markovitch 2005). This paper will develop an effective algorithm to approximate the solution of (2) for unsupervised feature selection.

### Identification Autoencoders (IAE)

To perform feature selection in the original space, our first attempt is to add a simple one-to-one layer between the input and hidden layers of AE to weigh the importance of each input feature. It is also natural to exploit the sparsity property of $l_1$ regularization for the weights of this layer for feature selection, inspired by Lasso (Tibshirani 1996). Then, we have the following formulation:

$$\min_{\mathbf{W}_{\mathrm{I}}, f, g} \|\mathbf{X} - f(g(\mathbf{X}\mathbf{W}_{\mathrm{I}}))\|_{\mathrm{F}}^2 + \lambda_1 \|\mathbf{W}_{\mathrm{I}}\|_1, \text{ s.t. } \mathbf{W}_{\mathrm{I}} \geqslant 0,$$

where $\mathbf{W}_{\mathrm{I}} = \mathrm{Diag}(\mathbf{w})$, $\mathbf{w} \in \mathbb{R}^m$, and $\lambda_1$ is a parameter balancing between the reconstruction error and sparsity regularization. The $\ell_1$ norm induces sparsity and shrinks the less important features' weights to 0, and it may make the features more discriminative as well. Here, we require that the entries of $\mathbf{W}_{\mathrm{I}}$ should be nonnegative since they represent the importance of the features and the non-negativity constraint would make their interpretation more meaningful (Xu et al. 2019). The fully connected concrete layer (Abid, Balin, and Zou 2019) has taken a similar non-negativity constraint, albeit for a full matrix.

For AE with such an additional one-to-one layer and the modified objective function, we call it an identification autoencoder (IAE) only for notational purpose. Actually, IAE is a general case of AgnoS-S (Doquet and Sebag 2019), where it does not impose any constraint on the dimension of the bottleneck layer of AE. After training, the features corresponding to the $k$ largest entries of $\mathbf{W}_\mathrm{I}$ are selected as the most informative features. Compared with standard AE, IAE clearly increases no more than $m$ additional parameters.

We may visualize the selected features by IAE in Figure 2 (c) and (e). It is seen that IAE captures a part of key features from the original samples; however, it cannot capture other key features on the skeleton of the digits, and the selected features fail to recover the original contents, as shown in Figure 2 (g). In general, the selected features by unsupervised feature selection are to be representative of the input data, implying that the selected features should reconstruct the original samples well. Thus, IAE cannot serve the purpose of feature selection in itself. Its failure is mainly due to the lack of diversity of its selected features. The $\ell_1$ regularization term in IAE may promote the sparsity of the feature weight vector; however, it cannot ensure a sufficient level of diversity needed by a representative subset of features. Because the features in real data often have significant inter-correlations and even redundancy, without properly taking account of them, the selected features would have high correlations yet lack necessary diversity. Directly computing the pairwise interactions of all features requires a full $m \times m$ weight matrix, which may be computationally costly for high-dimensional data. Accounting for higher-order interactions between features would require even higher complexities. To address this problem, we propose a simple yet effective approach by using a sub-NN to locally excavate for diversity information from the feature weights, thereby reducing the search space significantly. We will introduce this sub-network below, which leads to the architecture of FAE.

## Fractal Autoencoders (FAE)

To remedy the diversity issue of IAE, we further design a sub-NN term, which requires that the subset of $k$ selected features from $\mathbf{W}_\mathrm{I}$ should be so diverse as to still represent the global contents of original samples as much as possible. Putting together, our proposed model is as follows:

$$\min_{\mathbf{W}_\mathrm{I},f,g} \|\mathbf{X} - f(g(\mathbf{X}\mathbf{W}_\mathrm{I}))\|_\mathrm{F}^2 + \lambda_1\|\mathbf{X} - f(g(\mathbf{X}\mathbf{W}_\mathrm{I}^{\max_k}))\|_\mathrm{F}^2 \\ +\lambda_2\|\mathbf{W}_\mathrm{I}\|_1, \text{ s.t. } \mathbf{W}_\mathrm{I} \geqslant 0, \tag{3}$$

where $\mathbf{W}_\mathrm{I}^{\max_k} = \mathrm{Diag}(\mathbf{w}^{\max_k})$, and $\lambda_1$ and $\lambda_2$ are nonnegative balancing parameters. We call a NN corresponding to (3) fractal autoencoders (FAE), due to its seemingly self-similarity characteristic: A small proportion of features in the second term achieve a similar performance to the whole set of features in the first term for reconstructing the original data. This characteristic will be manifested more clearly when applying FAE to extract multiple feature subsets later.

In training, we solve $\mathbf{W}_\mathrm{I}^{\max_k}$ by jointly optimizing the global-NN and sub-NN. After training FAE, we obtain $\mathbf{W}_\mathrm{I}^{\max_k}$ which can be used to perform feature selection on new samples during testing. We illustrate the selected features, the selected features superimposed on the original samples (for easy visualization), and reconstructed samples with these features in (d), (f), and (h) of Figure 2, respectively, for 9 random samples from MNIST.

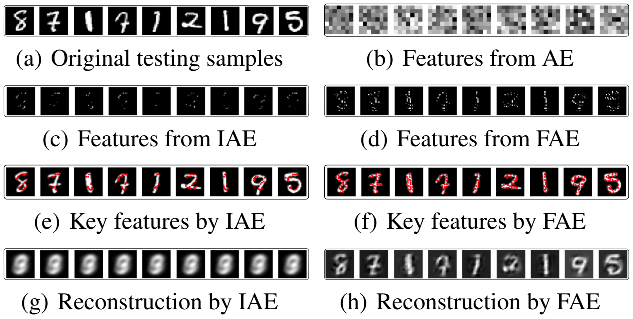

(a) Original testing samples
(b) Features from AE
(c) Features from IAE
(d) Features from FAE
(e) Key features by IAE
(f) Key features by FAE
(g) Reconstruction by IAE
(h) Reconstruction by FAE

Figure 2: (a) Testing samples randomly chosen from MNIST; (b) 49 features extracted by AE (size enlarged for visualization); (c) 50 features selected by IAE; (d) 50 features selected by FAE; (e) key features by IAE shown with original samples; (f) key features by FAE shown with original samples; (g) IAE's reconstruction based on the 50 features; (h) FAE's reconstruction based on the 50 features; (c)-(h) are best viewed with enlarging.

| | Dataset | # Sample | # Feature/# Gene | # Class |
|---|---|---|---|---|
| 1 | Mice Protein | 1,080 | 77 | 8 |
| 2 | COIL-20 | 1,440 | 400 | 20 |
| 3 | Activity | 5,744 | 561 | 6 |
| 4 | ISOLET | 7,797 | 617 | 26 |
| 5 | MNIST | 10,000 | 784 | 10 |
| 6 | MNIST-Fashion | 10,000 | 784 | 10 |
| 7 | USPS | 9,298 | 256 | 10 |
| 8 | GLIOMA | 50 | 4,434 | 4 |
| 9 | leukemia | 72 | 7,070 | 2 |
| 10 | pixraw10P | 100 | 10,000 | 10 |
| 11 | Prostate_GE | 102 | 5,966 | 2 |
| 12 | warpAR10P | 130 | 2,400 | 10 |
| 13 | SMK_CAN_187 | 187 | 19,993 | 2 |
| 14 | arcene | 200 | 10,000 | 2 |
| 15 | GEO | 111,009 | 10,463 | Null |

Table 1: Statistics of datasets.

# Experiments

In this section, we will perform experiments to extensively assess FAE by comparing it with contemporary methods on many benchmarking datasets.

## Datasets to Be Used

The benchmarking datasets used in this paper are Mice Protein Expression[1], COIL-20 (Nene, Nayar, and Murase 1996), Smartphone Dataset for Human Activity Recognition in Ambient Assisted Living (Anguita et al. 2013), ISOLET[2], MNIST (Lecun et al. 1998), MNIST-Fashion (Xiao, Rasul, and Vollgraf 2017), GEO[3], USPS, GLIOMA, leukemia, pixraw10P, Prostate_GE, warpAR10P, SMK_CAN_187, and arcene[4]. We summarize the statistics of these datasets in Table 1. Following CAE (Abid, Balin, and Zou 2019) and considering the long runtime of UDFS, for MNIST and MNIST-

[1] http://archive.ics.uci.edu/ml/datasets/Mice+Protein+Expression

[2] http://archive.ics.uci.edu/ml/datasets/ISOLET

[3] https://cbcl.ics.uci.edu/public_data/D-GEX

[4] The last eight datasets are from the scikit-feature feature selection repository (Li et al. 2017).

Fashion, we randomly choose $6,000$ samples from each training set to train and validate and $4,000$ from each testing set for testing. And we randomly split $6,000$ samples into training and validation sets at a ratio of $90:10$. For GEO, we randomly split the preprocessed GEO in the same way as D-GEX (Chen et al. 2016): $88,807$ for training, $11,101$ for validating, and $11,101$ for testing[5]. For other datasets, we randomly split them into training, validation, and testing sets by a ratio of $72:8:20$.

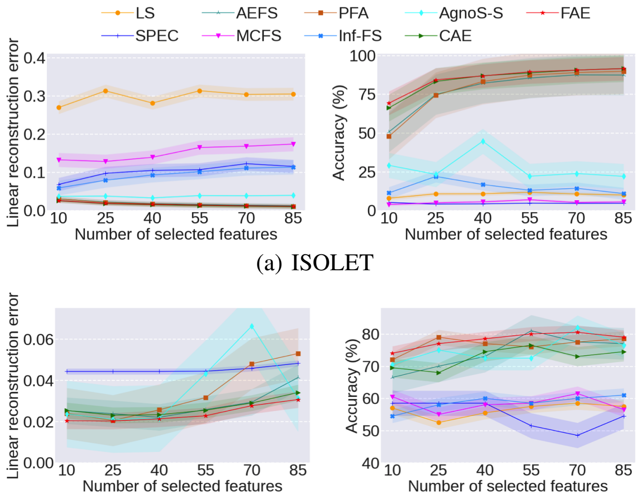
## Design of Experiments

In experiments of FAE, we set the maximum number of epochs to be $1,000$ for datasets 1-14 and 200 for dataset 15. We initialize the weights of feature selection layer by sampling uniformly from $\mathrm{U}[0.999999, 0.9999999]$ and the other layers with the Xavier normal initializer. We adopt the Adam optimizer (Kingma and Ba 2015) with an initialized learning rate of $0.001$. We set $\lambda_1$ and $\lambda_2$ in (3) to 2 and 0.1, respectively. For the hyper-parameter setting, we perform a grid search on the validation set, and then choose the optimal one. In the following experiments, we only use the linear version of FAE for simplicity, that is, $g(\mathbf{X}) = \mathbf{X}\mathbf{W}_\mathrm{E}$, $\mathbf{W}_\mathrm{E} \in \mathbb{R}^{m\times k}$, and $f(g(\mathbf{X})) = (g(\mathbf{X}))\mathbf{W}_\mathrm{D}$, $\mathbf{W}_\mathrm{D} \in \mathbb{R}^{k\times m}$. The simple, linear version of FAE can already achieve superior performance, as shown below.

For the specified number of selected features $k$, we adopt two options: 1) We take $k = 10$ for Mice Protein dataset, 50 for datasets 2-7 following CAE (Abid, Balin, and Zou 2019), and 64 for high-dimensional datasets 8-14. For all baseline methods, we adopt this option. 2) For FAE, we additionally use fewer features, with $k = 8$ for Mice Protein dataset, 36 for datasets 2-7, and 50 for datasets 8-14, to further show its superior representative ability over competing methods. We set the dimension of the latent space to $k$ and denote FAE with these two options as Opt1 and Opt2, respectively.

Two metrics are used for evaluating the models: 1) reconstruction error, which is measured in mean squared error (MSE); 2) classification accuracy, which is measured by passing the selected features to a downstream classifier as a viable means to benchmark the quality of the selected subset of features. For fair comparison, following CAE (Abid, Balin, and Zou 2019), after selecting the features, we train a linear regression model with no regularization to reconstruct the original features, and the resulting linear reconstruction error is used as the first metric[6]. Meanwhile, for the second metric we use the extremely randomized trees (Geurts, Ernst, and Wehenkel 2006) as the classifier.

All experiments are implemented with Python 3.7.8, Tensorflow 1.14, and Keras 2.2.5. The codes can be found at https://github.com/xinxingwu-uk/FAE.



(a) ISOLET

(b) arcene

Figure 3: Reconstruction and classification results versus $k$.

## Results on Fourteen Datasets

The experimental results on reconstruction and classification with the selected features by different algorithms are reported in Tables 2 and 3. For all the results, we implement 5 runs with random splits on the fixed dataset to present mean results and standard errors. From Table 2, it is seen that FAE yields smaller reconstruction errors than baseline methods on majority datasets, indicating its strong ability for representing the original data. From Table 3, it is evident that FAE exhibits consistently superior performance in the downstream classification task on most of the benchmarking datasets.

Further, we compare the behaviors of FAE with respect to $k$ with those of the baseline algorithms. By varying $k$ on ISOLET and arcene, we obtain the corresponding linear reconstruction errors and classification accuracies. We plot the results in Figure 3[7]. The results show that FAE performs better and more stable than other algorithms in most cases.

**Feature Importance** To examine the importance of the features selected by FAE, we rank and partition them into two equal groups, that is, each group has 25 features. The results are shown in Figure 4. We can observe that, in most cases, the classification accuracy of the first group is generally better than the second group. However, since FAE is unsupervised, some selected features that are essential for reconstruction might not be important for classification.

## Computational Complexity

Experimentally, the computational time of our algorithm (3) is about twice that of sparse AE. FAE has only an additional sub-NN compared to sparse AE and shares parameters with the global-NN. Also, the fitting error term of sub-NN is quadratic and similar to sparse AE's fitting error term. Thus,

[5] Abid, Balin, and Zou (2019) stated that they used the same preprocessing scheme with D-GEX. Though having the same number of features, we note that their dataset has a slightly different sample size $112,171$ from ours and that in (Chen et al. 2016).

[6] For reconstruction error only, it denotes the error from the second term of (3), that is, $\|\mathbf{X} - f(g(\mathbf{X}\mathbf{W}_\mathrm{I}^{\max_k}))\|_\mathrm{F}^2$.

[7] For better visualization, we ignore the algorithms with large linear reconstruction errors.

| Dataset | LS | SPEC | NDFS | AEFS | UDFS | MCFS | PFA | Inf-FS | AgnoS-S | CAE | FAE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | Opt1 | Opt2 |
| Mice Protein | .575±.118 | 1.32±1.78 | 1.69±0.98 | .020±.007 | **.009±.006** | 16.5±30.4 | .028±.002 | .443±.040 | .038±.014 | .032±.001 | .014±.005 | .015±.005 |
| COIL-20 | .225±.035 | .711±.626 | .144±.016 | .011±.0 | .015±.001 | 2.89±2.47 | **.009±.0** | .134±.013 | .035±.009 | .011±.001 | .011±.001 | .013±.001 |
| Activity | 4166±776 | .153±.048 | 284±137 | .005±.0 | **.004±.0** | 63.4±109.3 | .005±.0 | .207±.043 | .009±.001 | **.004±.0** | .005±.001 | .005±.001 |
| ISOLET | .304±.047 | .104±.007 | .144±.005 | .016±.0 | .019±.001 | .154±.032 | .015±.0 | .099±.015 | .035±.005 | **.013±.0** | .015±.0 | .017±.0 |
| MNIST | .305±.0 | .067±.001 | .134±.004 | .037±.002 | .029±.002 | .128±.003 | .028±.001 | .101±.003 | .055±.005 | **.019±.0** | **.019±.0** | .025±.001 |
| MNIST-Fashion | 11.4±22.5 | .109±.007 | .139±.010 | .023±.0 | .027±.003 | .458±.657 | .022±.0 | .105±.006 | .025±.001 | **.019±.0** | **.019±.0** | .022±.0 |
| USPS | 2.99±.73 | 1.28±.16 | 1.07±.08 | .027±.003 | .032±.002 | 1.07±.07 | .018±.002 | 4.05 ±.63 | .017±.003 | .012±.001 | **.011±.001** | .021±.001 |
| GLIOMA | .226±.033 | .259±.030 | .347±.044 | **.065±.008** | .072±.004 | .249±.019 | .067±.011 | .211±.065 | .068±.012 | .066±.011 | .069±.010 | .141±.032 |
| leukemia | 10.7±5.7 | 8.50±1.41 | 12.7±4.0 | 7.78±2.42 | \ | 14.1±4.0 | 6.30±1.22 | 12.3±2.7 | **6.14±1.07** | 398±736 | 7.01±.94 | 9.07±3.56 |
| pixraw10P | 31.1±29.8 | .554±.155 | .645±.400 | .006±.002 | \ | .163±.041 | .003±.001 | 1.357±.700 | .009±.004 | .013±.011 | .005±.004 | **.002±.001** |
| ProstateGE | 1.36±.42 | .404±.310 | 3.91±2.05 | .242±.094 | \ | 3.00±3.36 | .142±.039 | .273 ±.093 | .146±.026 | .202±.137 | .144±.039 | **.068±.016** |
| warpAR10P | 1.28±.61 | 4.73±3.95 | .597±.118 | .039±.007 | .086±.029 | 1.08±.33 | .036±.005 | 3.68±1.15 | .045±.006 | .074±.027 | .040±.005 | **.033±.005** |
| SMK_CAN_187 | 5.87±1.08 | .127±.020 | 3.52 ± .62 | .114±.022 | \ | 6.24±1.00 | .110±.016 | 6.84±2.53 | .102±.012 | .100±.015 | .105±.019 | **.097±.021** |
| arcene | .410±.250 | .045±.001 | 1493±267 | .055±.043 | \ | 1.86±.62 | .035±.009 | 478±205 | .030±.012 | .027±.001 | .025±.001 | **.023±.001** |

Table 2: Linear reconstruction error with selected features by different algorithms. The “\” mark denotes the case with prohibitive running time, where the algorithm ran for more than one week without getting the result and thus was stopped.

| Dataset | LS | SPEC | NDFS | AEFS | UDFS | MCFS | PFA | Inf-FS | AgnoS-S | CAE | FAE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | Opt1 | Opt2 |
| Mice Protein | 17.3±4.0 | 13.7±3.6 | 15.6±10.5 | 88.5±4.0 | **95.1±4.0** | 17.4±5.9 | 92.8±2.9 | 19.3±6.9 | 63.2±37.2 | 66.9±4.8 | 87.8±7.3 | 78.6±18.2 |
| COIL-20 | 16.0±3.4 | 16.4±2.6 | 16.8±3.9 | 99.3±.2 | 97.6±2.3 | 10.5±2.6 | 99.4±.4 | 34.4±9.0 | 83.5±14.5 | 98.8±.5 | **99.6±.3** | 99.0±1.0 |
| Activity | 29.0±1.4 | 20.3±1.1 | 18.2±1.2 | 88.7±1.7 | 91.7±1.5 | 21.6±4.6 | 88.8±1.4 | 24.1±5.0 | 74.9±10.7 | **91.9±1.0** | 91.4±1.1 | 88.8±1.3 |
| ISOLET | 11.1±.7 | 3.4±.9 | 8.4±1.6 | 82.9±2.1 | 73.9±5.8 | 6.0±1.2 | 86.5±1.3 | 14.0±1.9 | 36.2±17.0 | 87.9±.6 | **89.0±.6** | 87.0±1.3 |
| MNIST | 12.4±1.0 | 11.2±1.0 | 10.5±1.9 | 80.2±2.6 | 88.1±1.6 | 13.0±4.1 | 88.5±2.0 | 13.2±1.3 | 43.5±15.6 | 92.5±.4 | **92.9±.7** | 90.8±.9 |
| MNIST-Fashion | 17.1±2.8 | 27.2±1.7 | 15.2±4.5 | 79.4±1.5 | 79.3±1.2 | 16.3±.9 | 80.3±10.5 | 18.8±3.8 | 78.4±1.3 | 82.3±1.0 | **82.5±.6** | 80.8±.9 |
| USPS | 34.2±4.8 | 36.2±6.9 | 10.6±1.8 | 94.9±.7 | 94.5±.3 | 12.3±3.4 | 95.7±.3 | 18.2±4.2 | 95.6±.4 | 96.2±.4 | **96.3±.2** | 96.2±.3 |
| GLIOMA | 46.0±17.4 | 22.0±16.0 | 34.0±8.0 | 68.0±13.3 | 76.0±20.6 | 46.0±8.0 | 66.0±8.0 | 42.0±11.7 | 62.0±7.5 | 70.0±16.7 | **76.0±8.0** | 72.0±16.0 |
| leukemia | 52.0±12.2 | 58.7±8.8 | 57.3±13.7 | 76.0±12.4 | \ | 54.7±10.7 | 72.0±14.9 | 56.0±6.8 | 72.0±14.9 | 65.3±12.9 | 80.0±11.2 | **80.0±6.0** |
| pixraw10P | 51.0±8.6 | 9.0±5.8 | 19.0±8.0 | **100.0±0.0** | \ | 11.0±5.8 | **100.0±0.0** | 41.0±23.5 | **100.0±0.0** | 99.0±2.0 | **100.0±0.0** | **100.0±0.0** |
| ProstateGE | 43.8±9.2 | 55.2±13.7 | 56.2±17.4 | 85.7±8.5 | \ | 46.7±11.0 | 81.0±10.0 | 57.1 ±15.7 | 81.9±4.7 | 78.1±9.8 | 82.9±5.7 | **85.7±4.2** |
| warpAR10P | 11.5±5.4 | 7.7±4.9 | 12.3±7.1 | 76.9±5.4 | 64.6±8.9 | 15.4±5.4 | **82.3±6.7** | 10.8±3.8 | 71.5±7.1 | 70.0±10.7 | 71.5±3.1 | 63.8±7.9 |
| SMK_CAN_187 | 55.3±7.1 | 65.8±6.9 | 53.2 ± 14.6 | 61.1±4.5 | \ | 50.5±4.2 | 70.0±2.7 | 47.4±2.9 | 71.6±5.6 | 69.5±8.3 | **72.1±6.6** | 71.1±8.5 |
| arcene | 59.5±7.0 | 51.5±9.8 | 57.5±7.3 | 81.0±5.2 | \ | 56.5±4.6 | 80.0±5.7 | 59.5±5.3 | 80.0±4.5 | 74.5±8.9 | **84.0±6.6** | 80.0±4.5 |

Table 3: Classification accuracy (%) with selected features by different algorithms. The mark “\ ” is used similarly to Table 2.

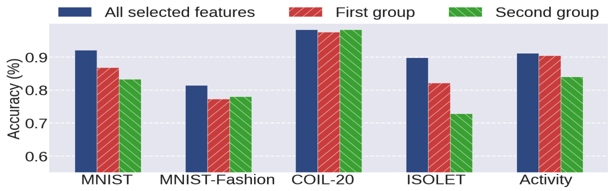
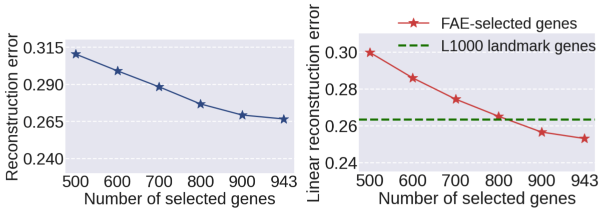


Figure 4: Analyses of feature importance with $k = 50$.



(a) (b)

Figure 5: Gene selection by using FAE for GEO. (a) Reconstruction error by FAE; (b) Reconstruction error by using the linear regression model on L1000 landmark genes and FAE-selected genes.

the overall computational complexity of FAE is of the same order as sparse AE.

## Analysis of L1000 Gene Expression

It is expensive to measure all gene expressions. To reduce the cost, researchers from the LINCS program[8] have found that a carefully selected set of genes can capture most gene expression information of the entire human transcriptome because the expression of genes is usually correlated under different conditions. Based on the selected genes, a linear regression model was used to infer the gene expression values of the remaining genes (Chen et al. 2016). Recently, Abid, Balin, and Zou (2019) have used CAE to further reduce the number of genes to 750 to achieve a similar linear reconstruction error about 0.3 to the original 943 landmark genes of L1000.

Now we apply FAE on the preprocessed GEO to select varying numbers of representative genes from 500 to 943. Figure 5 (a) shows that, by using 600 genes, FAE achieves a reconstruction error better than that with 750 selected genes by CAE. However, CAE uses a slightly different number of samples with ours. For consistency, we mainly compare FAE with L1000. We compute the reconstruction error by using the linear regression model on the genes selected by FAE and the landmark genes of L1000, and the results in MSE are depicted in Figure 5 (b). Evidently, using 800 genes by FAE achieves a similar reconstruction to L1000. Thus, FAE reduces the number of genes by about 15% compared to L1000. The selected genes by FAE are displayed in Figure 6. It is observed that, with different numbers of selected genes, a few genes sometimes are selected and sometimes not, which may be attributed to the significant correlations among genes. In addition, when selecting the same number of 943 genes, only 90 genes selected by CAE are among the landmark genes, while 121 by FAE are among the landmark genes. These results indicate that L1000 landmark genes can

[8] See http://www.lincsproject.org/LINCS/

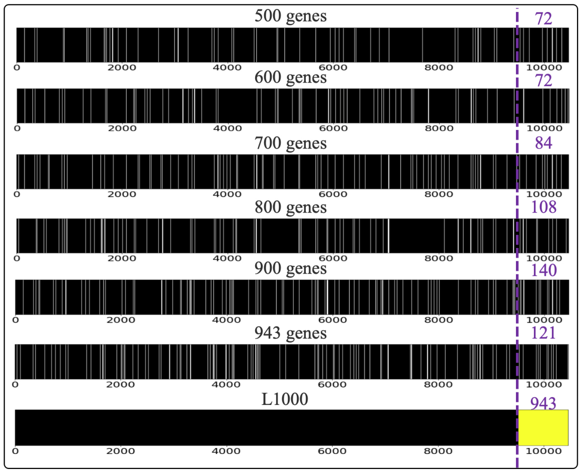


Figure 6: Comparison of different numbers of selected genes with 943 landmark gene. The white lines denote those selected genes by FAE. The purple dashed line separates 943 landmark genes (color coded in yellow) from the other genes. The purple numbers 72, 72, 84, 108, 140, and 121 denote respectively the numbers of overlapping genes between the landmark genes and those selected by FAE with $k$ being 500, 600, 700, 800, 900, and 943.

be significantly enhanced in representation power.

## An Application of FAE

FAE is applicable and easily extensible to different tasks. Here we show an application of exploiting multiple hierarchical subsets of the key features.

### $h$-HFAE

For an image, usually there are many pixels highly correlated with each other. Thus, the subsets of key features might not be unique; indeed, there often exists more than one subset of informative features that can recover the original data well. Excavating these potential subsets of meaningful features is conductive to facilitate data compression (Sousa et al. 2007) and better understand the structure and interrelationship of the features. Yet, almost all existing feature selection approaches have little ability to explore these potential subsets. To achieve such an ability, we develop an application in the framework of FAE, which selects multiple non-overlapping subsets of representative features. For clarity, we formalize it as follows:

$$\min_{\mathbf{W}_{\mathrm{I}}^{\max_{k,i}},\mathbf{W}_{\mathrm{E}},\mathbf{W}_{\mathrm{D}}} \|\mathbf{X}-((\mathbf{X}\mathbf{W}_{\mathrm{I}})\mathbf{W}_{\mathrm{E}})\mathbf{W}_{\mathrm{D}}\|_{\mathrm{F}}^2+\lambda_0\|\mathbf{W}_{\mathrm{I}}\|_1 \\ +\sum_{i=1}^{h}\lambda_i\|\mathbf{X}-((\mathbf{X}\mathbf{W}_{\mathrm{I}}^{\max_{k,i}})\mathbf{W}_{\mathrm{E}})\mathbf{W}_{\mathrm{D}}\|_{\mathrm{F}}^2,\ \text{s.t. } \mathbf{W}_{\mathrm{I}}\geqslant 0, \tag{4}$$

where $\mathbf{W}_{\mathrm{I}}^{\max_{k,1}} = \mathrm{Diag}(\mathbf{w}^{\max_{k,1}})$, $\mathbf{W}_{\mathrm{I}}^{\max_{k,i}} = \mathrm{Diag}((\mathbf{w}/\mathbf{w}^{\max_{k,i-1}})^{\max_{k,i}})$, $i=2,\ldots,h$, $h$ is the number of desired subsets of relevant features, $\lambda_i, i=0,\ldots,h$, are hyper-parameters, and $(\mathbf{w}/\mathbf{w}^{\max_{k,i-1}})^{\max_{k,i}}$ is an operation to retain the $i$-th group of $k$ largest entries from $\mathbf{w}$ while making zero all the other entries including the $(i-1)$ groups of $k$ largest entries of $\mathbf{w}^{\max_{k,1}}$, $\mathbf{w}^{\max_{k,2}}$, . . ., and $\mathbf{w}^{\max_{k,i-1}}$. In (4), the first two terms estimate the importance of each input feature globally; then, the remaining terms organize the top $kh$ features into $h$ hierarchical subsets in descending order of importance values, with each subset having $k$ features. These $h$ subsets are selected by using $h$ sub-NNs, which work together in an orchestrated way: The $(i+1)$-th sub-NN exploits the $(i+1)$-th hierarchical subset of features by leaving out the $i$ subsets of features found by the previous $i$ sub-NN(s). For notational convenience, we denote this application for identifying multiple hierarchical subsets of features by $h$-HFAE.

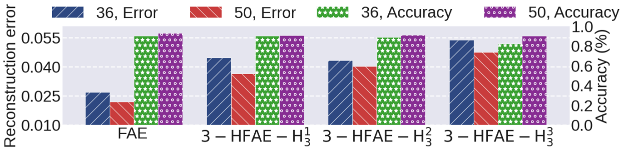


Figure 7: Reconstruction and classification results of FAE and 3-HFAE on MNIST. 3-HFAE-$\mathrm{H}_3^1$, 3-HFAE-$\mathrm{H}_3^2$, and 3-HFAE-$\mathrm{H}_3^3$ denote respectively the first three hierarchical subsets of selected features.

To verify the effectiveness of $h$-HFAE, we set $h = 3$ and apply it to MNIST. The reconstruction and classification results together with those of FAE are shown in Figure 7, where we set the hyper-parameters $\lambda_0$, $\lambda_1$, $\lambda_2$, and $\lambda_3$ in (4) to be $0.05$, $1.5$, $2$, and $3$, respectively. With 50 selected features per group, different hierarchies of 3-HFAE achieve almost the same accuracy; with 36 selected features per group, the third group of features from 3-HFAE-$\mathrm{H}_3^3$ has slightly worse accuracy than the other two groups. This result implies that 50 selected features per group are more stable for 3-HFAE. For reconstruction error, the vanilla version of FAE is the best among all results.

CAE (Abid, Balin, and Zou 2019) displays the relationships of the top 3 selected features at each node of the concrete selector layer; however, the second and third top features might be insignificant due to the potentially trivial average probability ($\leqslant 0.01$). Different from CAE, $h$-HFAE uses the weights to assign the features into different hierarchical subsets for selection and exploration. In the Supplementary Material, we demonstrate that for 3-HFAE there exists a considerable degree of similarity between different hierarchical subsets of selected features. Thus, $h$-HFAE can reveal the redundancy or high correlations among features.

## Conclusions

In this paper, we propose a new framework for unsupervised feature selection, which extends AE by adding a simple one-to-one layer and a sub-NN to achieve both global exploring of representative abilities of the features and local mining for their diversity. Extensive assessment of the new framework has been performed on real datasets. Experimental results demonstrate its superior performance over contemporary methods. Moreover, this new framework is applicable and easily extensible to other tasks and we will further extend it in our future work.

## Acknowledgments

This work is supported in part by NSF OIA2040665, NIH R56NS117587, and R01HD101508. We sincerely thank the anonymous reviewers for their valuable comments.

## References

Abid, A.; Balin, M. F.; and Zou, J. 2019. Concrete autoencoders: differentiable feature selection and reconstruction. In *International Conference on Machine Learning*, 444–453. Long Beach, California, United States.

Alelyani, S.; Tang, J.; and Liu, H. 2013. Feature selection for clustering: a review. In Aggarwal, C. C.; and Reddy, C. K., eds., *Data Clustering: Algorithms and Applications*, chapter 2, 29–60. CRC Press, 1st edition.

Anafi, R. C.; Francey, L. J.; Hogenesch, J. B.; and Kim, J. 2017. CYCLOPS reveals human transcriptional rhythms in health and disease. *The Proceedings of the National Academy of Sciences* 114(20): 5312–5317.

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; and Reyes-Ortiz, J. L. 2013. A public domain dataset for human activity recognition using smartphones. In *European Symposium on Artificial Neural Networks, Computational Intelligence And Machine Learning*, 437–442. Bruges, Belgium.

Ballard, D. H. 1987. Modular learning in neural networks. In *National Conference on Artificial Intelligence*, 279–284. Seattle, Washington, United States.

Bellman, R. 1957. *Dynamic programming*. Princeton, New Jersey, United States: Princeton University Press, 1st edition.

Blum, A.; Hopcroft, J.; and Kannan, R. 2020. *Foundations of data science*. Cambridge, United Kingdom: Cambridge University Press, 1st edition.

Cai, D.; Zhang, C.; and He, X. 2010. Unsupervised feature selection for multi-cluster data. In *International conference on Knowledge discovery and data mining*, 333–342. Washington, District of Columbia, United States.

Chen, Y.; Li, Y.; Narayan, R.; Subramanian, A.; and Xie, X. 2016. Gene expression inference with deep learning. *Bioinformatics* 32(12): 1832–1839.

Cheng, Q.; Zhou, H.; and Cheng, J. 2011. The Fisher-Markov selector: fast selecting maximally separable feature subset for multiclass classification with applications to high-dimensional data. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(6): 1217–1233.

Clarke, R.; Ressom, H. W.; Wang, A.; Xuan, J.; Liu, M. C.; Gehan, E. A.; and Wang, Y. 2008. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer* 8(1): 37–49.

Doquet, G.; and Sebag, M. 2019. Agnostic feature selection. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 343–358. Würzburg, Germany.

Geurts, P.; Ernst, D.; and Wehenkel, L. 2006. Extremely randomized trees. *Machine learning* 63(1): 3–42.

Guyon, I.; and Elisseeff, A. 2003. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3(7-8): 1157–1182.

Hamo, Y.; and Markovitch, S. 2005. The COMPSET algorithm for subset selection. In *International Joint Conference on Artificial Intelligence*, 728–733. Edinburgh, Scotland, United Kingdom.

Han, K.; Wang, Y.; Zhang, C.; Li, C.; and Xu, C. 2018. Autoencoder inspired unsupervised feature selection. In *International Conference on Acoustics, Speech and Signal Processing*, 2941–2945. Calgary, Alberta, Canada.

He, X.; Cai, D.; and Niyogi, P. 2005. Laplacian score for feature selection. In *Advances in Neural Information Processing Systems*, 507–514. Vancouver, British Columbia, Canada.

Ji, P.; Zhang, T.; Li, H.; Salzmann, M.; and Reid, I. 2017. Deep subspace clustering networks. In *Advances in Neural Information Processing Systems*, 23–32. Long Beach, California, United States.

Kingma, D. P.; and Ba, J. L. 2015. Adam: a method for stochastic optimization. In *International Conference for Learning Representations*. San Diego, California, USA.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. *arXiv:1312.6114v10, https://arxiv.org/abs/1312.6114* .

Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.

Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R. P.; Tang, J.; and Liu, H. 2017. Feature selection: a data perspective. *ACM Computing Surveys* 50(6): 94.

Li, Z.; Yang, Y.; Liu, J.; Zhou, X.; and Lu, H. 2012. Unsupervised feature selection using nonnegative spectral analysis. In *AAAI Conference on Artificial Intelligence*, 1026–1032. Toronto, Ontario, Canada.

Lu, Y.; Cohen, I.; Zhou, X. S.; and Tian, Q. 2007. Feature selection using principal feature analysis. In *International conference on Multimedia*, 301–304. Augsburg, Bavaria, Germany.

Maddison, C. J.; Mnih, A.; and Teh, Y. W. 2017. The concrete distribution: a continuous relaxation of discrete random variables. *arXiv: 1611.00712v3, https://arxiv.org/abs/1611.00712* .

Natarajan, B. K. 1995. Sparse approximate solutions to linear systems. *SIAM Journal on Computing* 24(2): 227–234.

Nene, S. A.; Nayar, S. K.; and Murase, H. 1996. Columbia object image library (COIL-20). Technical Report CUCS-005-96, Department of Computer Science, Columbia University, New York, United States.

Pearson, K. 1901. 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* 2(11): 559–572.

Peng, C.; Kang, Z.; Hu, Y.; Cheng, J.; and Cheng, Q. 2017. Nonnegative matrix factorization with integrated graph and feature learning. *ACM Transactions on Intelligent Systems and Technology* 8(3): 1–29.

Peng, C.; Kang, Z.; Yang, M.; and Cheng, Q. 2016. Feature selection embedded subspace clustering. *IEEE Signal Processing Letters* 23(7): 1018 –1022.

Roffo, G.; Melzi, S.; and Cristani, M. 2015. Infinite feature selection. In *International Conference on Computer Vision*, 4202–4210. Santiago, Chile.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1985. Learning internal representations by error propagation. ICS Report 8506, Institute for Cognitive Science, University of California, San Diego, La Jolla, California, United States.

Sousa, C. M.; Cavalcante, A. B.; Guilhon, D.; and Barros, A. K. 2007. Image compression by redundancy reduction. In *International conference on Independent component analysis and signal separation*, 422–429. London, United Kingdom.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1): 267–288.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv: 1708.07747v2, https://arxiv.org/pdf/1708.07747.pdf* .

Xie, J.; Girshick, R.; and Farhadi, A. 2016. Unsupervised deep embedding for clustering analysis. In *International Conference on Machine Learning*, 478–487. New York, United States.

Xu, J.; Yu, M.; Shao, L.; Zuo, W.; Meng, D.; Zhang, L.; and Zhang, D. 2019. Scaled simplex representation for subspace clustering. *IEEE Transactions on Cybernetics* 1–13.

Yang, Y.; Shen, H. T.; Ma, Z.; Huang, Z.; and Zhou, X. 2011. $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning. In *International Joint Conference on Artificial Intelligence*, 1589–1594. Barcelona, Catalonia, Spain.

Zhao, Z.; and Liu, H. 2007. Spectral feature selection for supervised and unsupervised learning. In *International Conference on Machine Learning*, 1151–1157. Corvallis, Oregon, United States.