EBEK: Exemplar-Based Kernel Preserving Embedding

Ahmed Elbagoury, Rania Ibrahim, Mohamed S. Kamel, and Fakhri Karray
University of Waterloo, Waterloo, Ontario, Canada N2L 3G1
{ahmed.elbagoury, rania.ibrahim, mkamel and karray}@uwaterloo.ca

Abstract

With the rapid increase in the available data, it becomes computationally harder to extract useful information. Thus, several techniques like PCA were proposed to embed high-dimensional data into low-dimensional latent space. However, these techniques don't take the data relations into account. This motivated the development of other techniques like MDS and LLE which preserve the relations between the data instances. Nonetheless, all these techniques still use latent features, which are difficult for data analysts to understand and grasp the information encoded in them. In this work, a new embedding technique is proposed to mitigate the previous problems by projecting the data to a space described by few points (i.e, exemplars) which preserves the relations between the data points. The proposed method **E**xemplar-**b**as**e**d **K**ernel Preserving (EBEK) embedding is shown theoretically to achieve the lowest reconstruction error of the kernel matrix. Using EBEK in approximate nearest neighbor task shows its ability to outperform related work by up to 60% in the recall while maintaining a good running time. In addition, our interpretability experiments show that EBEK's selected basis are more understandable than the latent basis in images datasets.

1 Introduction

Recently, extracting useful information from large volume of data has attracted many researchers in different areas like text, images, videos and more. Nonetheless, this large size of data has very high computational and memory demands, in addition it is hard for data analysts to have a grasp of large sized data that has high dimensions. Therefore, dimensionality reduction techniques are used as a solution to these challenges. Some of these techniques, like Principal Component Analysis (PCA) [Jolliffe, 2002], summarize the data by projecting it on some latent space. However, these latent features are difficult to interpret. Using exemplar-based embedding solves the aforementioned problem by projecting the data into a lower dimension space spanned by a subset of data points (i.e., the exemplars), which attains lucid features, that are related to explicit data points. In addition, these exemplars can be used by the data analysts to gain a better understanding of the data nature and structure.

One criterion for selecting the exemplars is minimizing the discrepancy between the original data matrix and the low rank approximation obtained by these exemplars, which is a combinatorial problem. Thus, many techniques have been proposed to solve it greedily as in [Farahat *et al.*, 2013]. One limitation of these techniques is not taking the data points relations and similarities into account, preserving such relations is shown to be effective in the similarity preserving dimensionality reduction techniques like Multidimensional Scaling (MDS) [Silva and Tenenbaum, 2002] and Locally Linear Embedding (LLE) [Roweis and Saul, 2000]. It is shown to be effective in topic detection [Elbagoury *et al.*, 2015] and in clustering [Ibrahim *et al.*, 2016]. Additionally, preserving the pairwise similarities in the embedded data is very useful for the task of Approximate Nearest Neighbor (ANN) search, which is defined as finding the set of samples that have the smallest distance to a given query sample. Finding ANNs has a wide range of applications in machine learning and information retrieval [Moraleda, 2008]. That is why in this work, we propose Exemplar-based Kernel Preserving (EBEK) embedding to choose the exemplars and the embedding that result in the best low rank approximation of the similarities of the data where the similarities are represented by a kernel matrix. In addition, formulating the problem as preserving the data similarities obviates the need to solve a combinatorial problem as will be shown in our theoretical analysis.

It is essential to develop techniques that can work on kernel matrices, as not all types of data can be represented in numerical feature vectors form. For instance, there is a need to group users in social media based on their friendship relations and to group proteins in bioinformatics based on their structures [Elgohary *et al.*, 2013]. EBEK supports linear kernel matrices and can be extended to support other kernels types. The contributions of the paper can be summarized as follows:

- Derive theoretical proof to show that Exemplar-based Linear Kernel Preserving embedding achieves the minimum reconstruction error for the kernel matrix.
- Evaluate the proposed approach in practical domains like approximate nearest neighbors search.
- Show the interpretability of the exemplars chosen by the

proposed approach on images datasets.

The rest of the paper is organized as follows: Section 2 gives an overview of related work techniques, then section 3 shows the details of the proposed Exemplar-based Kernel Preserving embedding. At the end, experimental evaluations are shown in section 4 and section 5 concludes our work.

2 Related Work

In this section, we shed the light on some of the embedding techniques that have been proposed in the literature. Then, we will discuss some of the popular techniques for finding the approximate nearst neighbors.

Several dimensionality reduction techniques project the data to some latent space to optimize a certain objective function, such as PCA [Jolliffe, 2002], which has two main disadvantages: 1) The produced basis are latent and do not have a clear meaning. 2) PCA doesn't consider data points relations and structure which, in some domains, are essential to be preserved in the low-dimensional space. Therefore, several dimensionality reduction techniques have been proposed to preserve data points relations like MDS [Silva and Tenenbaum, 2002] and LLE [Roweis and Saul, 2000]. For instance, classical MDS [Silva and Tenenbaum, 2002] minimizes the difference between the Euclidean distances of the data points in the original space and the Euclidean distances of the projected data points in the lower dimension space and hence maintains the relationships between the points. LLE [Roweis and Saul, 2000] is another dimensionality reduction technique that aims to find a mapping, which preserves the local distances between the points, by trying to reconstruct the points only using their k-nearest neighbors. On the other hand, our approach combines both preserving the data points relations and structures, while representing the new basis using samples that can be easily interpreted by data analysts.

Computing ANNs is a time and memory consuming task to be performed on the high-dimensional data. That is why, several approaches have been proposed to project the data into a lower dimension space and then utilize this lower dimension space to compute the nearest neighbors of the data points like in [Gong and Lazebnik, 2011], [Andoni and Indyk, 2006] and [Raginsky and Lazebnik, 2009]. For example, a modified version of PCA called PCA-RR is used in [Gong and Lazebnik, 2011] where the projection matrix W of PCA is multiplied by a random orthogonal matrix R and then the approach uses WR as lower dimension basis to project the data on. It was also shown in [Gong and Lazebnik, 2011] that PCA-RR outperforms PCA in ANNs task. Yet the objective function of PCA does not preserve the similarities of the data, which limits its ability to find the best ANNs. Local Sensitivity Hashing (LSH) [Andoni and Indyk, 2006] mitigates this problem by trying to preserve the local neighbors of the points using random hash functions that with high probability map similar data points to the same buckets. While, utilizing these buckets enables LSH to retrieve the ANNs, defining general random hash functions for LSH is a difficult task. SKLSH [Raginsky and Lazebnik, 2009] modifies the objective function of LSH to approximate shift-invariant kernels using random feature mapping. ITQ [Gong and Lazebnik, 2011] is another approach for finding ANNs, which tries to learn a similarity preserving binary coding using training data and then utilizes it to encode the data and compute the ANNs. While ITQ coding captures the data properties, it requires a lot of training data to find a good binary coding, in addition this training phase consumes a lot of time.

3 EBEK: Exemplar-based Kernel Preserving Embedding

3.1 Notations

The following notations are used throughout the rest of the paper unless otherwise is stated. Scalars are denoted by small letters (e.g., m, n), sets are shown in script letters (e.g., $\mathcal{E}, \mathcal{H}$), vectors are denoted by small bold italic letters (e.g., $\boldsymbol{f}, \boldsymbol{g}$), and matrices are denoted by capital letters (e.g., A, S). In addition the following notations are used:

For a set $\mathcal{E}$:	
$\lvert\mathcal{E}\rvert$	the size of the set.
For a vector $\boldsymbol{x} \in \mathbb{R}^{\mathrm{m}}$:	
$\boldsymbol{x}_i$	i-th element of $\boldsymbol{x}$.
For a matrix $A \in \mathbb{R}^{\mathrm{n}\times\mathrm{m}}$:	
$A_{i,j}$	the (i,j)-th entry of A.
$A_{i,:}$	the i-th row of A.
$A_{:,j}$	the j-th column of A.
$A_{:,\mathcal{E}}$	the submatrix of A which consists of the set $\mathcal{E}$ of columns.
$A_{\eta,\mathcal{E}}$	the submatrix of A that consists of the set η of rows and the set $\mathcal{E}$ of columns.
A^T	the transpose of A.
$\tilde{A}$	the low rank approximation of A.
$\lvert\lvert A\rvert\rvert_F$	the Frobenius norm of A.

We start this section by providing the details of embedding that preserves linear kernels in subsection 3.2 and then we explain the details of extending the approach to support any arbitrary kernels.

3.2 Exemplar-based Linear Kernel Preserving Embedding

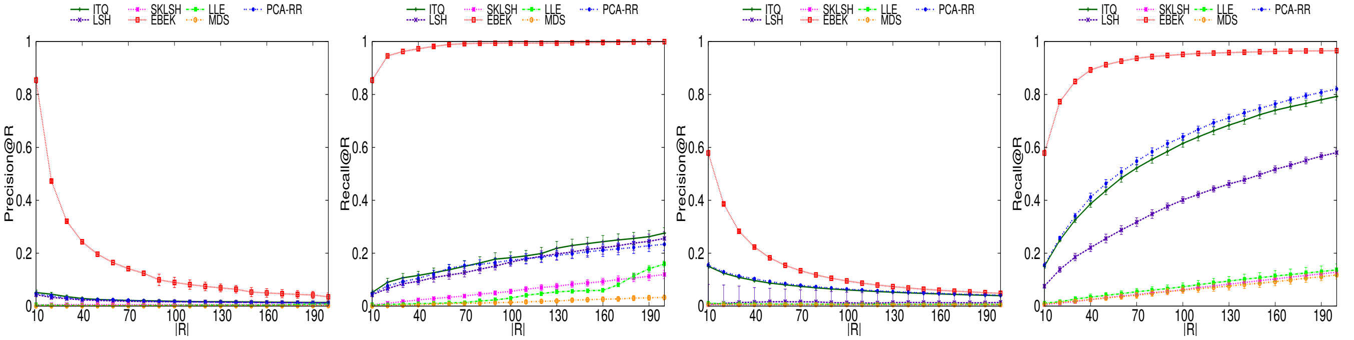
Our objective in this work is choosing a subset of columns (i.e, data points) that preserve the pairwise similarities as much as possible between the data embedded in the span of these columns. In addition, we would like these columns to be less similar to each other to ensure that these columns capture the different characteristics of the dataset. Given this objective the problem can be defined as follows.

Problem Definition Given a data matrix $A \in \mathbb{R}^{d\times n}$ (n samples in d dimensional space). Select a subset $\mathcal{E}$ of m columns, such that:

$$\arg\min_{A_{:,\mathcal{E}},T} ||S - \tilde{S}||_F = \arg\min_{A_{:,\mathcal{E}},T} ||A^T A - \tilde{A}^T \tilde{A}||_F$$

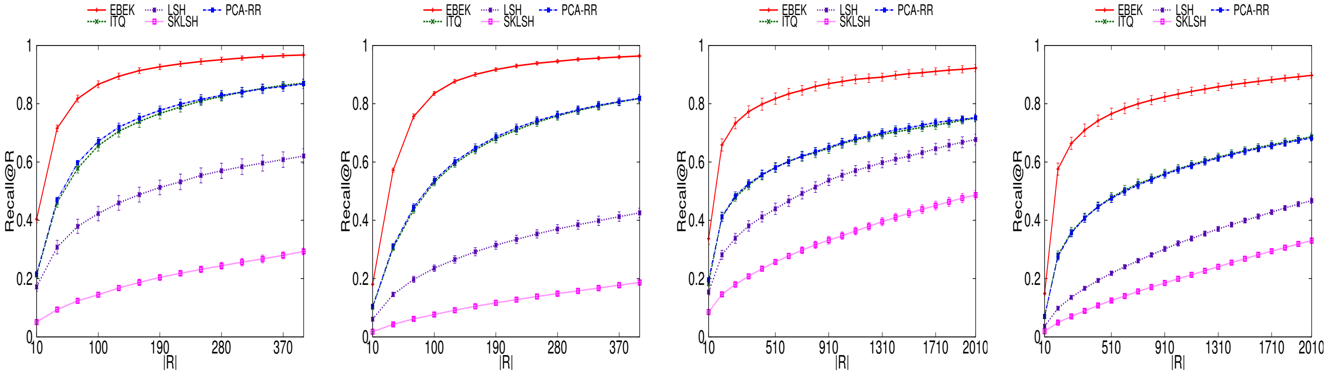
$$\text{s.t. } S(i,j) \leq \epsilon \quad \forall i,j \in \mathcal{E} \tag{1}$$

Where $\tilde{A}$ is the low rank approximation of A using the columns $A_{:,\mathcal{E}}$, $\tilde{A} = A_{:,\mathcal{E}}T$, $A_{:,\mathcal{E}} \in \mathbb{R}^{d\times m}$ and $T \in \mathbb{R}^{m\times n}$. T



(a) COIL20 Dataset Precision@R (b) COIL20 Dataset Recall@R (c) ISOLET Dataset Precision@R (d) ISOLET Dataset Recall@R

Figure 1: Precision@R and Recall@R for COIL20 and ISOLET Datasets



(a) TDT2 Dataset, $|\mathcal{T}| = 10$ (b) TDT2 Dataset, $|\mathcal{T}| = 50$ (c) 20NG Dataset, $|\mathcal{T}| = 10$ (d) 20NG Dataset, $|\mathcal{T}| = 50$

Figure 2: Recall@R for TDT2 and 20NG Datasets

represents the coefficients used to reconstruct the n samples using the m selected samples and ϵ is a similarity threshold to ensure that the columns are not similar to each other.

At first we will drop the constraints on the columns similarities and try to minimize the objective function and then we will show how to minimize the objective function while preserving these constrains. The goal of this objective function is to choose $A_{:,\mathcal{E}}$ and T that minimize the Frobenius norm of the difference between the pairwise similarity matrix of the original data ($S = A^T A$) and the pairwise similarity matrix of the low rank approximation data ($\tilde{S} = \tilde{A}^T \tilde{A}$), where the similarity here is defined by the linear kernel. Therefore, the problem can be reduced to:

$$\arg\min_{A_{:,\mathcal{E}},T} ||A^T A - \tilde{A}^T \tilde{A}||_F = \tag{2}$$

$$\arg\min_{A_{:,\mathcal{E}},T} ||A^T A - T^T A^T_{:,\mathcal{E}} A_{:,\mathcal{E}} T||_F$$

Let $S_{\mathcal{E},\mathcal{E}} = A^T_{:,\mathcal{E}} A_{:,\mathcal{E}} \in \mathbb{R}^{m \times m}$, which represents the pairwise similarities between the selected m samples. Then, equation 2 can be rewritten as:

$$\arg\min_{A_{:,\mathcal{E}},T} ||S - T^T S_{\mathcal{E},\mathcal{E}} T||_F \tag{3}$$

As the matrix S is symmetric positive semi-definite (by construction), then $S = V\Sigma^2 V^T$, where $A = U\Sigma V^T$, is the singular value decomposition of A. In addition, Σ is a diagonal matrix with $\texttt{rank}(S)$ positive elements and $n - \texttt{rank}(S)$ zero elements on the diagonal. Then equation 3 can be written as:

$$\arg\min_{A_{:,\mathcal{E}},T} ||V\Sigma^2 V^T - T^T S_{\mathcal{E},\mathcal{E}} T||_F \tag{4}$$

Lemma 3.1. *$||GBQ||_F = ||B||_F$ for any matrix B and orthogonal matrices G and Q.*

Proof.

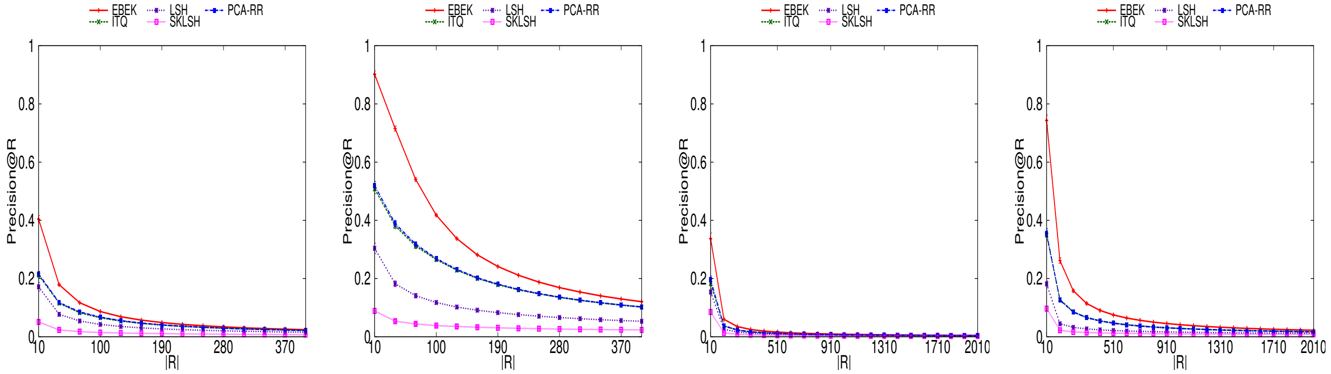
$$||GBQ||_F^2 = tr((GBQ)^T GBQ)$$
$$= tr(Q^T B^T G^T GBQ)$$
$$= tr(QQ^T B^T G^T GB) = tr(B^T B) = ||B||_F$$

As Q and G are orthogonal matrices, hence $QQ^T = I$ and $G^T G = I$.

□

Based on lemma 3.1, equation 4 can be re-written as:

$$\arg\min_{A_{:,\mathcal{E}},T} ||V^T (V\Sigma^2 V^T - T^T S_{\mathcal{E},\mathcal{E}} T) V||_F$$
$$= \arg\min_{A_{:,\mathcal{E}},T} ||\Sigma^2 - V^T T^T S_{\mathcal{E},\mathcal{E}} TV||_F$$
$$= \arg\min_{A_{:,\mathcal{E}},T} ||\Sigma^2 - (TV)^T S_{\mathcal{E},\mathcal{E}} TV||_F \tag{5}$$



(a) TDT2 Dataset, $|\mathcal{T}| = 10$ (b) TDT2 Dataset, $|\mathcal{T}| = 50$ (c) 20NG Dataset, $|\mathcal{T}| = 10$ (d) 20NG Dataset, $|\mathcal{T}| = 50$

Figure 3: Precision@R for TDT2 and 20NG Datasets

Lemma 3.2. *$S_{\mathcal{E},\mathcal{E}} = P^{-T}DP^{-1}$, where $P \in \mathbb{R}^{m\times m}$ is an invertible matrix and $D \in \mathbb{R}^{m\times m}$ is a diagonal matrix that has only entries 0 and $+1$. The number of $+1$ in D equals r, where r is the rank of matrix $S_{\mathcal{E},\mathcal{E}}$.*

Proof. Using Sylvester's Law of Inertia [Sylvester, 1852], each symmetric matrix $E \in \mathbb{R}^{m\times m}$ is congruent to a diagonal matrix $D \in \mathbb{R}^{m\times m}$ which has only entries 0, $+1$ and -1 along the diagonal, where the number of zero diagonal elements is $m-p$, $p = \text{rank}(E)$, the number of positive diagonal elements, q, is the number of positive eigenvalues, the number of negative diagonal elements is the number of negative eigenvalues $p-q$. Which means that there exists an invertible matrix $P \in \mathbb{R}^{m\times m}$ such that: $P^TEP = D$. Applying this to the matrix $S_{\mathcal{E},\mathcal{E}}$ gives the following: $P^TS_{\mathcal{E},\mathcal{E}}P = D$, then

$$S_{\mathcal{E},\mathcal{E}} = P^{-T}DP^{-1} \tag{6}$$

As the matrix $S_{\mathcal{E},\mathcal{E}}$ is symmetric positive semi-definite, then it has r positive eigenvalues, where $r = \text{rank}(S_{\mathcal{E},\mathcal{E}}) = \text{rank}(A_{:,\mathcal{E}})$, and $m-r$ zero eigenvalues. The matrix P can be obtained by multiplying pairs of elementary transformations, one of which is with rows and the other is the corresponding transformation with the columns as explained in [Lipschutz and Lipson, 2012]. □

Theorem 3.3. *By setting $T = P(\Sigma_{1:m,:})V^T$, where P satisfies equation 6 and selecting a subset $\mathcal{E}$ of columns from matrix A that have the highest rank, the matrix $\tilde{S}$, which equals to $T^TA^T_{:,\mathcal{E}}A_{:,\mathcal{E}}T$, achieves the minimum low rank approximation of S.*

Proof. Using lemma 3.2, and by substituting equation 6 in equation 5, we get:

$$\arg\min_{A_{:,\mathcal{E}},T} ||\Sigma^2 - (TV)^TP^{-T}DP^{-1}TV||_F$$

$$= \arg\min_{A_{:,\mathcal{E}},T} ||\Sigma^2 - (P^{-1}TV)^TDP^{-1}TV||_F \tag{7}$$

Our objective is to put the matrix D in canonical form such that:

$$D = \begin{bmatrix} I_r & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \tag{8}$$

Where I_r is r-by-r identity matrix. As the singular values of S are sorted along the diagonal of Σ, putting the matrix D in the form of equation 8 enables us to cancel the first r singular values (the largest ones), which means the error of equation 7 in terms of the Frobenius norm will be $\sqrt{\sum_{i=r+1}^{n}\sigma_i^4}$, where σ_i^2 is the i^{th} singular value of S. This can be achieved by setting the value $P^{-1}TV = (\Sigma_{1:m,:})$, where $\Sigma_{1:m,:}$ is the first m rows of the matrix Σ. This can be seen by substituting the value of $P^{-1}TV$ in equation 7, which will be:

$$\arg\min_{A_{:,\mathcal{E}},T} ||\Sigma^2 - (\Sigma_{1:m,:})^TD(\Sigma_{1:m,:})||_F \tag{9}$$

Thus, the optimal value for $P^{-1}TV = (\Sigma_{1:m,:})$, and $T = P(\Sigma_{1:m,:})V^T$. The error in this case is equal to the minimum achieved error using rank-k approximation obtained by SVD [Zhang *et al.*, 2002]. So to minimize 7, we need to:

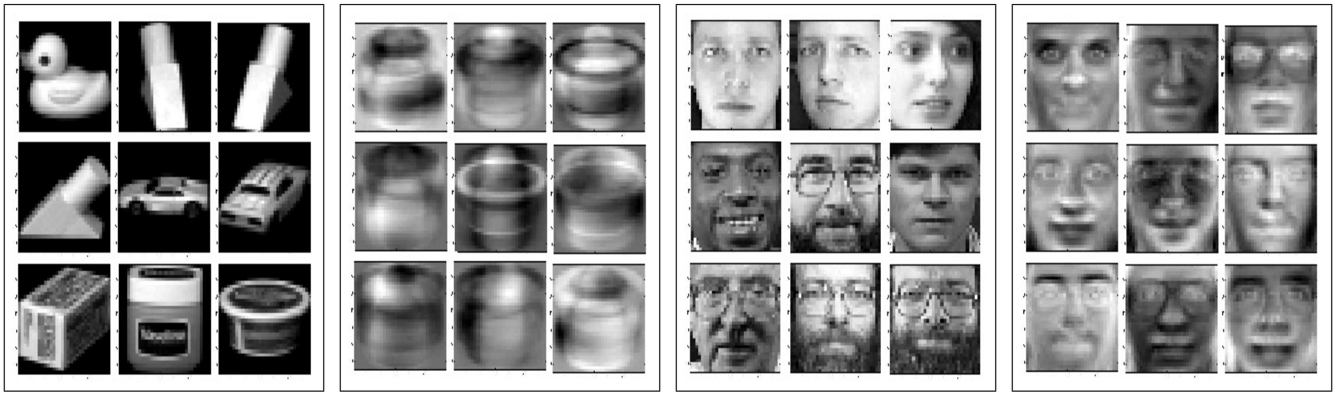
- Choose subset of columns $\mathcal{E}$ from A that have the maximum rank.
- Set the value of T to $P(\Sigma_{1:m,:})V^T$.

□

To maximize the rank of $A_{:,\mathcal{E}}$, there are two cases:

- If $r \geq m$, we can choose any independent m columns by reducing the matrix to its echelon column form and use the non-zero columns.
- If $r < m$, in this case we will use the non-zero columns of the echelon form and any other $m-r$ columns, and in this case the error will be zero.

Recall that $\tilde{A} = A_{:,\mathcal{E}}T$. To obtain the lower dimension embedding of the data in the space spanned by $A_{:,\mathcal{E}}$ we will replace the matrix $A_{:,\mathcal{E}}$ by its QR factorization. $\tilde{A} = QRT$, where Q is the orthogonal bases of the space spanned by $A_{:,\mathcal{E}}$ and the lower dimension embedding of the data is RT. Until



(a) EBEK Basis in COIL20 Dataset (b) PCA Basis in COIL20 Dataset (c) EBEK Basis in ORL Dataset (d) PCA Basis in ORL Dataset

Figure 4: EBEK and PCA Basis in COIL20 and ORL Datasets

now, we have only considered minimizing the objective function without the similarities constrains. As shown in the previous proof, choosing any subset of columns that has the maximum rank will be optimal for the objective function. Therefore, to decide which subset to choose, we employ the similarities constrains and choose a subset of columns with the maximum rank, such that the pairwise similarities between these columns are upper-bounded by a similarity threshold ϵ that can be chosen empirically.

Algorithm 1 shows the pseudo code of the algorithm. The method `getIndependentcol` returns m independent columns. This can be computed using echelon form. However, the set of independent columns can be computed more efficiently using algorithm 2. The algorithm starts with an arbitrary column, as the first independent column (line 2), then for each subsequent column j check if it has component orthogonal to the previously chosen columns or not (lines 5 to 7); if it has and its similarity with each of the previously selected columns is less than ϵ , then it will be included in the set of independent columns (lines 8 to 10).

EBEK running time complexity is $O(dn \log m + (d + n)m^2 + nmd + m^3)$, where $O(dn \log m + (d + n)m^2)$ is the time to compute the stochastic SVD decomposition of A, $O(nmd)$ for the independent columns selection and $O(m^3)$ for computing the matrix P. Note that, QR decomposition complexity is dominated by the other steps as the QR decomposision is done only for the selected columns in $O(dm^2)$. To extend EBEK to work on any arbitrary kernel K, we get the SVD of the K instead of A and then we get the independent columns from the kernel matrix and use the rest of the approach to embed the kernel matrix into the selected lower-dimensional space.

4 Experimental Results

The effectiveness of the proposed approach is evaluated on two tasks, approximate nearest neighbors search and interpretability experiments. Section 4.1 shows the setup and results for the ANNs search task, and section 4.2 shows the interpretability experiments.

Algorithm 1: Linear Kernel Preserving Embedding

```
Data: Matrix $A \in \mathbb{R}^{d \times n}$ and integer $m$
Result: $W \in \mathbb{R}^{m \times n}$, which represents the lower
        dimension embedding of the data
1 $[\Sigma, V] \leftarrow$ stochasticSVD($A$, $m$)
2 $\mathcal{E} \leftarrow$ getIndependentcol($A, m$)
3 $S_{\mathcal{E},\mathcal{E}} \leftarrow A^T_{:,\mathcal{E}} A_{:,\mathcal{E}}$
4 $P \leftarrow$ diagMatrix($S_{\mathcal{E},\mathcal{E}}$)
5 $T \leftarrow P \Sigma V^T$
6 $[Q, R] =$ orthogonalize($A_{:,\mathcal{E}}$)
7 $W \leftarrow RT$
```

4.1 Approximate Nearest Neighbors Search

In this subsection, we discuss the experimental setup and results for the task of ANNs search. To evaluate the effectiveness of our approach, the pairwise similarities between the lower dimension data is computed and the nearest neighbors are retrieved based on the lower dimension embedding.

To build the ground truth, the set $\mathcal{T}$ of nearest neighbors is retrieved by computing the distance to all queries and then applying the linear scan. The search quality for each approach is measured using Recall@$\mathcal{R}$ and Precision@$\mathcal{R}$ as in [Zhang *et al.*, 2014], where for each query the set of $\mathcal{R}$ nearest neighbors is retrieved and the recall is computed as the fraction of the samples in both set $\mathcal{T}$ and $\mathcal{R}$ and the size of the set $\mathcal{T}$, Recall@$\mathcal{R} = \frac{|\mathcal{R} \cap \mathcal{T}|}{|\mathcal{T}|}$ and Precision@$\mathcal{R} = \frac{|\mathcal{R} \cap \mathcal{T}|}{|\mathcal{R}|}$.

We have used four datasets, COIL20 which contains 1440 samples in 1024 dimensional space, ISOLET which contains 1560 samples in 617 dimensions, TDT2 which contains 9394 sample in 19677 dimensional space and a subset of 20 Newsgroups (20NG in short) containing 9990 samples in 29360 dimensional space [Cai *et al.*, 2009]. The performance with $|\mathcal{T}| = 10$, and 50 is reported and each experiment is repeated 10 times and the average and 95% confidence interval are provided. The observed behavior remains valid for other $|\mathcal{T}|$. Additionally, the number of the basis in the lower dimensional space m is set to 10 in all the techniques. Note that the results in this subsection are not affected by the value of ϵ

Algorithm 2: getIndependentcol: Independent Columns Selection

```
Data: Matrix $A \in \mathbb{R}^{d \times n}$, integer $m$, $\epsilon$
Result: $\mathcal{E}$ a set of indexes of $m$ independent columns in
        matrix $A$
size $\leftarrow 1; \mathcal{E} \leftarrow \{1\}$
for $i = 2$ : $min(m, n)$ do
    $a_i \leftarrow A_{:,i}$
    for $j = 1$ : $size$ do
        $a_j \leftarrow A_{:,\mathcal{E}(j)}$
        $a_i \leftarrow a_i - \frac{<a_i,a_j>}{<a_j,a_j>} a_j$
    if $||a_i||_1 \neq 0$ and $S(A_{:,i}, A_{:,\mathcal{E}}) \leq \epsilon$ then
        size $\leftarrow$ size $+ 1$
        $\mathcal{E} \leftarrow \mathcal{E} \cup i$
```

Algorithm 3: diagMatrix: Diagonalize the Input Matrix

```
Data: Matrix $S \in \mathbb{R}^{m \times m}$
Result: $P$
$I \in \mathbb{R}^{m \times m}$ identity matrix
for $i = 2 : m$ do
    $D_{i:} = S_{i:} / \sqrt{|S_{i,i}|}$
    $D_{:i} = S_{:i} / \sqrt{|S_{i,i}|}$
    $I_{i:} = I_{i:} / \sqrt{|S_{i,i}|}$
    for $j = i + 1 : m$ do
        mult $= -\frac{D_{j,i}}{D_{i,i}}$
        $D_{j:} =$ mult $\times D_{i:} + D_{j:}$
        $D_{:j} =$ mult $\times D_{:i} + D_{:j}$
        $I_{j:} =$ mult $\times I_{i:} + I_{j:}$
    $P = I^T$
```

as discussed in section 3.2. Figure 1 shows the results of the different techniques in COIL20 and ISOLET datasets and as shown in the figure, EBEK was able to achieve the best Precision@R and Recall@R. After that, ITQ and PCA-RR were the second best in Precision@R and Recall@R. Moreover, table 1 shows the running time of the techniques in COIL20 and ISOLET datasets. The results show that LSH and SKLSH were the fastest approaches, while EBEK was the third fastest approach with a gap of at most 0.04 seconds to LSH.

Figures 2 and 3 show the effect of changing the $|\mathcal{T}|$ on the Recall@$\mathcal{R}$ and Precision@$\mathcal{R}$ for both TDT2 and 20NG datasets. Note that, MDS and LLE were omitted from TDT2 and 20NG datasets as they were taking more than 20 minutes to run. Table 2 shows the running time of obtaining the low dimension embedding and the bit-encoding (depending on the approach) in TDT2 and 20NG datasets. It is obvious that EBEK consistently achieves the highest precision and recall while achieving the lowest running time.

Table 1: Approximate Nearest Neighbors Running Time (in Seconds) Comparison for COIL20 and ISOLET Datasets

	COIL20	**ISOLET**
EBEK	0.07 ± 0.00	0.03 ± 0.01
ITQ	0.08 ± 0.00	0.05 ± 0.00
LSH	**0.03 ± 0.00**	**0.02 ± 0.00**
PCA-RR	0.07 ± 0.00	0.04 ± 0.00
SKLSH	**0.03 ± 0.00**	**0.02 ± 0.00**
MDS	129.39 ± 0.20	146.44 ± 1.23
LLE	0.43 ± 0.01	1.06 ± 0.05

Table 2: Approximate Nearest Neighbors Running Time (in Seconds) Comparison for TDT2 and 20NG Datasets

	TDT2 $|\mathcal{T}| = 10$	**TDT2** $|\mathcal{T}| = 50$	**20NG** $|\mathcal{T}| = 10$	**20NG** $|\mathcal{T}| = 50$
EBEK	**5.17 ± 0.07**	**5.17 ± 0.08**	**6.21 ± 0.07**	**6.35 ± 0.1**
ITQ	32.77 ± 0.96	32.19 ± 1.66	260.40 ± 40.80	335.27 ± 74.82
LSH	14.47 ± 0.11	14.27 ± 0.18	268.09 ± 91.93	276.29 ± 80.68
PCA-RR	31.87 ± 1.12	31.70 ± 1.25	300.09 ± 58.29	451.66 ± 91.01
SKLSH	14.58 ± 0.27	14.16 ± 0.17	868.07 ± 130.50	716.33 ± 198.58

4.2 Interpretability Experiments

In order to show that the basis detected by EBEK are more understandable than the basis detected by the other approaches, two datasets are selected which are COIL20 and ORL datasets and the basis selected by EBEK and PCA-RR are drawn in figure 4. Note that the basis of PCA were similar to the basis detected by PCA-RR and PCA-RR has much better quality in the approximate nearest neighbor task, thus we only show PCA-RR basis. The value of ϵ was chosen empirically to yield the best visualization results and was set to 0.65 and 0.94 in COIL20 and ORL datasets respectively. As shown in the figure 4, in COIL20 dataset, EBEK basis were more interpretable than PCA basis, as it shows that COIL20 contains different objects in different orientations. While PCA produced understandable basis in the ORL dataset, still EBEK basis are more understandable. PCA basis point out that there is a change in the mouth area in the dataset images and as a viewer you do not know what are these changes. However, EBEK shows you these changes with men with beards and people with different mouth emotions. Additionally, EBEK basis capture characteristics that PCA can not capture, for example that the dataset contains different gender, different age and different color people.

5 Conclusion

In this paper, **E**xemplar-**b**as**e**d **K**ernel Preserving (EBEK) embedding is proposed and shown theoretically to achieve the lowest reconstruction error of the kernel matrix. Evaluation results show that EBEK exceeds the related work in the retrieved Approximate Nearest Neighbors (ANNs) quality, while maintaining a good running time. Moreover, our interpretability experiments show that EBEK's selected basis are more understandable than the latent basis of the images datasets.

References

[Andoni and Indyk, 2006] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 459–468. IEEE, 2006.

[Cai *et al.*, 2009] Deng Cai, Xuanhui Wang, and Xiaofei He. Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the 26th Annual International Conference on Machine Learning (ICML'09)*, pages 105–112, 2009.

[Elbagoury *et al.*, 2015] Ahmed Elbagoury, Rania Ibrahim, Ahmed K Farahat, Mohamed S Kamel, and Fakhri Karray. Exemplar-based topic detection in twitter streams. In *Ninth International AAAI Conference on Web and Social Media*, 2015.

[Elgohary *et al.*, 2013] Ahmed Elgohary, Ahmed K Farahat, Mohamed S Kamel, and Fakhri Karray. Embed and conquer: scalable embeddings for kernel k-means on mapreduce. *CoRR abs/1311.2334*, 2013.

[Farahat *et al.*, 2013] Ahmed K Farahat, Ahmed Elgohary, Ali Ghodsi, and Mohamed S Kamel. Distributed column subset selection on mapreduce. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 171–180. IEEE, 2013.

[Gong and Lazebnik, 2011] Yunchao Gong and Svetlana Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 817–824. IEEE, 2011.

[Ibrahim *et al.*, 2016] Rania Ibrahim, Ahmed Elbagoury, Mohamed S Kamel, and Fakhri Karray. Lvc: Local variance-based clustering. In *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016.

[Jolliffe, 2002] IT Jolliffe. Principal components as a small number of interpretable variables: some examples. *Principal Component Analysis*, pages 63–77, 2002.

[Lipschutz and Lipson, 2012] Seymour Lipschutz and Marc Lipson. *Schaum's Outline of Linear Algebra*. McGraw-Hill Education, 5 edition, 2012.

[Moraleda, 2008] Jorge Moraleda. Gregory shakhnarovich, trevor darrell and piotr indyk: Nearest-neighbors methods in learning and vision. theory and practice. *Pattern Analysis and Applications*, 11(2):221–222, 2008.

[Raginsky and Lazebnik, 2009] Maxim Raginsky and Svetlana Lazebnik. Locality-sensitive binary codes from shift-invariant kernels. In *Advances in neural information processing systems*, pages 1509–1517, 2009.

[Roweis and Saul, 2000] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[Silva and Tenenbaum, 2002] Vin D Silva and Joshua B Tenenbaum. Global versus local methods in nonlinear dimensionality reduction. In *Advances in neural information processing systems*, pages 705–712, 2002.

[Sylvester, 1852] James Joseph Sylvester. A demonstration of the theorem that every homogeneous quadratic polynomial is reducible by real orthogonal substitutions to the form of a sum of positive and negative squares. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 4(23):138–142, 1852.

[Zhang *et al.*, 2002] Zhenyue Zhang, Hongyuan Zha, and Horst Simon. Low-rank approximations with sparse factors i: Basic algorithms and error analysis. *SIAM Journal on Matrix Analysis and Applications*, 23(3):706–727, 2002.

[Zhang *et al.*, 2014] Ting Zhang, Chao Du, and Jingdong Wang. Composite quantization for approximate nearest neighbor search. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 838–846, 2014.