2019 IEEE Symposium Series on Computational Intelligence (SSCI)
December 6-9 2019, Xiamen, China

Principal Component Analysis on Graph-Hessian

Yichen Pan
College of Control Science and Engineering
China University of Petroleum (East China)
Qingdao, China
Email: panyc_upc@163.com

Weifeng Liu
College of Control Science and Engineering
China University of Petroleum (East China)
Qingdao, China
Email: liuwf@upc.edu.cn

Yicong Zhou
Faculty of Science and Technology
University of Macau
Macau, China
Email: yicongzhou@um.edu.mo

Liqiang Nie
School of Computer Science and Technology
Shandong University
Qingdao, China
Email: nieliqiang@gmail.com

Abstract—Principal Component Analysis (PCA) is a widely used linear dimensionality reduction method, which assumes that the data are drawn from a low-dimensional affine subspace of a high-dimensional space. However, it only uses the feature information of the samples. By exploiting structural information of data and embedding it into the PCA framework, the local positional relationship between samples in the original space can be preserved, so that the performance of downstream tasks based on PCA can be improved. In this paper, we introduce Hessian regularization into PCA and propose a new model called Graph-Hessian Principal Component Analysis (GHPCA). Hessian can correctly use the intrinsic local geometry of the data manifold. It is better able to maintain the neighborhood relationship between data in high-dimensional space. Compared with other Laplacian-based models, our model can obtain more abundant structural information after dimensionality reduction, and it can better restore low-dimensional structures. By comparing with several methods of PCA, GLPCA, RPCA and RPCAG, through the K-means clustering experiments on USPS handwritten digital dataset, YALE face dataset and COIL20 object image dataset, it is proved that our models are superior to other principal component analysis models in clustering tasks.

Keywords—*dimensionality reduction; principal component analysis; manifold learning; graph; hessian regularization*

I. INTRODUCTION

With the advent of the era of Big Data[1], the size, dimensionality, and complexity of data are expanding at an explosive rate. In some high-dimensional datasets, the number of features is much larger than the sample size, making it impossible to learn valuable information through them. In addition, the data is easy to introduce useless information in the process of acquisition[2], causing it often contains significant amounts of noise, corrupted entries or outliers. Therefore, how to extract useful information from destroyed high-dimensional data and reduce the data dimensionality has become one of the main tasks of machine learning.

At present, the dimensionality reduction has many applications in the fields such as pattern recognition and computer vision. In order to accomplish this task, different types of dimensionality reduction algorithms are introduced. Principal Component Analysis[2][3] is one of the most classical dimensionality reduction algorithm, which projects data into low-dimensional space by the orthogonal transformation. However, it has drawbacks in some respects, so many extended algorithms based on PCA are proposed. These algorithms can be broadly divided into five categories: robust PCA methods, sparse PCA methods, probabilistic PCA methods, kernel PCA methods and manifold regularized PCA methods.

A. Robust Principal Component Analysis Methods

Due to the existence of quadratic terms, PCA is more sensitive to outliers. In response to this problem, Ding et al.[4] proposed Rotational Invariant L1-norm Principal Component Analysis (R1-PCA), which effectively handles outliers by proposing an R1 norm constraint. Candes et al.[5] proposed a new solution to make PCA more robust to outliers, named Robust Principal Component Analysis (RPCA). It decomposes the original data matrix into a sum of a low-rank matrix and a sparse matrix through a penalty term, and the corrupted data is decomposed into a sparse matrix.

B. Sparse Principal Component Analysis Methods

The principal components (PCs) obtained by PCA is a linear combination of all the original variables. So it is often difficult to interpret PCs. Zou et al. [6]used the lasso (elastic net) to propose a new method called Sparse Principal Component Analysis (SPCA), which produces a modified principal component with sparse loadings. Seghouan et al.[7] introduced Adaptive Block Sparse Principal Component Analysis (BSPCA) to generate improved principal components with block sparse loads. Shen et al. [8]proposed a sparse Principal Component Analysis via regularized singular value decomposition (sPCA-rSVD), using the connection of PCA with SVD and extracting the PCs by solving the low-rank matrix approximation.

C. Probabilistic Principal Component Analysis Methods

PCA does not consider the probability distribution of the data and lacks the relevant probability model of the observed

data. Tipping et al. [9]introduced a new method called Probabilistic Principal Component Analysis (PPCA) to determine the principal axis of data from the maximum likelihood estimation in a Gaussian latent variable model which is related to the factor analysis. Lawrence et al. [10]proposed the dual probabilistic Principal Component Analysis (DPPCA) so that linear mapping from the embedded space can be easily nonlinearized by Gaussian processes.

D. Kernel Principal Component Analysis Methods

In many real-world tasks, the proper low-dimensional embedding can be found through nonlinear mapping. To solve this issue, Schölkopf et al. [11]used the kernel function to nuclearize PCA and proposed Kernel Principal Component Analysis (KPCA). Ding et al.[12]proposed an Adaptive Kernel Principal Component Analysis (AKPCA) method, which can flexibly and accurately track kernel principal components to overcome the batch nature of KPCA and adaptively adjust kernel principal components.

E. Manifold Regularized Principal Component Analysis Methods

For a given dataset, we can use its feature and structure information. And we can build graph[13], i.e., structure information between the samples by the feature data. Therefore, the difference between vector data and graph is sometimes ambiguous. We can use multiple data sources to get better results. Based on this idea, Zhang et al.[14] proposed Manifold Regularization Low-Rank Matrix Factorization (MMF), which incorporates manifold regularization to the matrix factorization. Jiang et al. [15]embed graph into the classical PCA and proposed Graph-Laplacian Principal Component Analysis (GLPCA), which combines the classic PCA with Laplacian embedding. The structural relationships between the samples are embedded in a low-dimensional representation. In order to enhance the robustness of the algorithm to outliers while introducing the graph information, Shahid et al.[16] proposed Robust Principal Component Analysis on Graphs (RPCAG), which embeds Laplacian embedding into the RPCA.

Because Laplacian seeks the problem of the first derivative, the solution is biased towards constant and the extrapolation ability is not strong, and the obtained structural information is not rich enough. To solve these problems, we propose to introduce Hessian regularization[17][18][19][20] into the framework of principal component analysis. Hessian has a richer null space, and due to the geodesic function in null space, it can correctly utilize the intrinsic local geometry of the data manifold, correctly reflect the positional relationship between the samples, and obtain richer structural information, so that the low-dimensional representation contains more abundant graph information. In this paper, K-means clustering experiments were performed in handwritten digital database USPS, face database YALE and object image database COIL20. Compared with several dimensionality reduction methods of PCA, GLPCA, RPCA and RPCAG, it is proved that our model is superior to other models in clustering tasks.

***Expansion*:** We also introduce Hessian into the framework of RPCA to enhance the robustness of the algorithm to outliers, to verify the superiority of the Hessian-based models and the Laplacian-based models, and the generalization ability of Hessian.

The rest of this article is organized as follows. Section Ⅱ provides some related works. In Section Ⅲ, we present our proposed GHPCA and GHRPCA algorithms. The solution of the algorithms is given in Section Ⅳ. Experimental setup and experimental results are given in Section Ⅴ. Finally, we conclude this article in Section Ⅵ.

II. RELATED WORK

In this section, we review several related works involved in our model, including principal component analysis, robust principal component analysis, and then analyze Hessian regularization and Laplacian regularization to prove the advantages of Hessian in maintaining local structure information.

A. Principal Component Analysis

Given a data matrix $X = [x_1,x_2,\cdots,x_n] \in R^{p\times n}$ containing n p -dimensional sample data, the classical PCA finds k orthogonal feature vectors, and forms a k-dimensional linear subspace $U \in R^{p\times d}$ .A projection matrix $Y \in R^{d\times n}$ is obtained by projecting X into the U. Principal component analysis finds U and Y by minimizing:

$$\min_{U,Y} \|X - UY\|_F^2 \ s.t.\ U^TU = I \tag{1}$$

where $U = [u_1,u_2,\cdots,u_d]$ is called the principal direction, and $Y = [y_1,y_2,\cdots,y_n]$ is the principal component of our demand. $\|\cdot\|_F$ is the Fourier norm of matrix. Using the obtained principal component Y instead of the original data matrix X to perform other tasks.

B. Robust Principal Component Analysis

When performing tasks, the data used is generally mixed with noise, which will affect the output of the task. A potential assumption of the classical PCA is that the noise involved is Gaussian, but less robust to other types of noise. Robust PCA proposes to use the form of matrix decomposition to eliminate the influence of noise on it. Its model is shown in:

$$\min_{A,S} \|A\|_* + \lambda\|S\|_1 \ s.t.\ X = A + S \tag{2}$$

where $A \in R^{p\times n}$ is the low-rank matrix containing the required information, which is the product of U and Y obtained from the classical PCA($A = UY$). $S \in R^{p\times n}$ is the sparse matrix containing noise information, and the parameters λ control the proportion of the sparse matrix S. $\|\cdot\|_*$ is the nuclear norm of matrix, $\|\cdot\|_1$ is the l_1 norm.

C. Laplacian Regularization and Hessian Regularization

Suppose $C^\infty(M)$ is the smooth functions set on the manifold M,M in the Euclidean space. Laplacian regularization as follows:

$$S_\Delta(f) = \int_M \|\nabla f\|^2 dV(x) \tag{3}$$

where $f \in C^{\infty}(M)$ and $f : M \to R$, $dV(x)$ represents a volume element, and S_{Δ} is the null space which $\{f \in C^{\infty}(M) | S(f) = 0\}$, is just constant function on M. $S : C^{\infty}(M) \to R$ represents the regularization.

Similarly, Hessian regularization is defined as:

$$S_{Hess}(f) = \int_M \sum_{r,s=1}^{m} \left(\frac{\partial^2 f}{\partial x_r \partial x_s}\right)^2 dV(x) \qquad (4)$$

Hessian regularization performs a second-order partial derivative on f. Eells and Lemaire[21] proved the following proposition.

Proposition 1. (Eells and Lemaire[21]) A function $f : M \to R$ with $f \in C^{\infty}(M)$ has zero second derivative, $\nabla_a \nabla_b f|_x = 0, \forall x \in M$, if and only if for any geodesic $\gamma : (-\varepsilon, \varepsilon) \to M$ parameterized by arc length s, there exists a constant c_{γ} depending on γ such that

$$\frac{\partial}{\partial s} f(\gamma(s)) = c_{\gamma}, \forall -\varepsilon < s < \varepsilon \qquad (5)$$

These functions which satisfy $\frac{\partial}{\partial s} f(\gamma(s)) = c_{\gamma}$ are called geodetic functions. They reflect constant changes of geodesic distance in the manifold. Therefore, the Hessian regularization's null space are linear functions with a constant change in geodesic distance, in other words, the null space of Hessian is more abundant than Laplacian. And because of the geodesic functions in null space, the Hessian regularization can better maintain the local structure.

III. Graph-Hessian Principal Component Analysis

In this part, we will introduce the principal component analysis on Graph-Hessian, and its extended model robust principal component analysis on Graph-Hessian. There is a data matrix containing n p-dimensional sample data, we will introduce our proposed models GHPCA and GHRPCA.

A. Graph-Hessian Principal Component Analysis

For Principal Component Analysis, we want to make the low-dimensional representation to include richer graph information. So we build Hessian regularization into the classic PCA framework and propose the GHPCA model:

$$\min_{U,Y} \|X - UY\|_F^2 + \gamma tr(YHY^T) \ s.t. \ YY^T = I \qquad (6)$$

here $U \in R^{p \times d}$ is the principal direction, $Y \in R^{d \times n}$ is the principal component. $\|\cdot\|_F$ is the Fourier norm of matrix, the parameter $\gamma \geq 0$, and it controls the smoothness of Y and the proportion of the two parts in the model, and H is the Hessian matrix, $tr(\cdot)$ represents the trace of the matrix.

The first half of the model represents the reconstruction error of the classical PCA, ensuring that the resulting low-dimensional representation reflects the original data to the greatest extent. The second half is Hessian regularization. Using Y and Hessian to construct regularization, the principal component Y can benefit from the sample similarity graph. The two-part combination can make Y benefit both at the same time. And this model has a closed-form solution, the solution can be solved by the Lagrange multiplier method, which will be given in the next section.

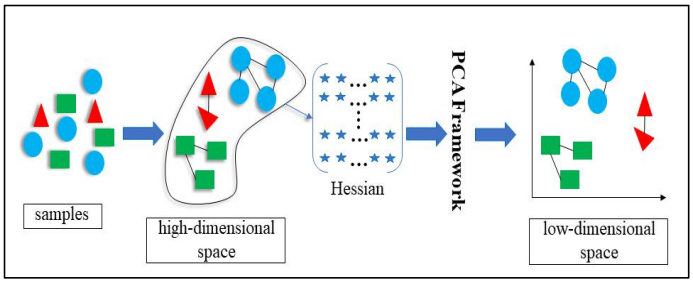

Fig. 1. The framework of GHPCA.

B. Graph-Hessian Robust Principal Component Analysis

For the Robust Principal Component Analysis, we hope that the low-rank matrix A contains the manifold information between the samples, so the matrix A and the Hessian matrix are used to form the regularization embedded into the framework of the RPCA. The GHRPCA model as follows:

$$\min_{A,S} \|A\|_* + \lambda \|S\|_1 + \gamma tr(AHA^T) \ s.t. \ X = A + S \qquad (7)$$

here A is the low-rank matrix, S is the sparse matrix, H is the Hessian matrix and the parameters control the sparsity of the model, and parameters $\gamma \geq 0$ control the smoothness of A. $\|\cdot\|_*$ is the nuclear norm of matrix, $tr(\cdot)$ represents the trace of the matrix, $\|\cdot\|_1$ is the l_1 norm.

The first two parts of the model represent the decomposition model of RPCA, resulting in low-rank matrix A containing useful information, separating the noise S. The latter part is Hessian regularization. Using A and Hessian to construct regularization, the graph information can be embedded in the low-rank matrix A. A is the product of U and Y, so that both matrices can be affected by the graph and reflect the original data more realistically. We use the ADMM algorithm to solve the model, as shown in section Ⅳ.

Laplacian is a constant function along the null space of the underlying manifold, while Hessian has a richer null space and contains richer local geometric relationships. Therefore, the obtained principal component Yor low-rank matrix A contains more abundant structural information. Due to the existence of second-order partial derivatives, Hessian has a stronger ability to construct sample neighbor relations, which makes the relationship between similar samples more closely when performing downstream tasks such as clustering.

IV. Algorithm

In this section, we present solutions of the above two models separately, and briefly introduce the solution steps of the Hessian matrix and analyze the computational complexity of our models.

A. Graph-Hessian Principal Component Analysis

For the GHPCA model, it has a closed-form solution, and the solution process is as follows:

Using the Lagrange multiplier method for (6) to get:

$$\min_{U,Y} f = \min_{U,Y} \|X - UY\|_F^2 + \gamma tr(YHY^T) + \mu(I - YY^T) \qquad (8)$$

Algorithm 1 Closed solution of GHPCA

1: Input X, H, γ

2: Calculating matrix $(-X^TX+\gamma H)$

3: Calculating the Covariance matrix $(-X^TX+\gamma H)(-X^TX+\gamma H)^T$

4: Calculating the eigenvector $[y_1,y_2,\cdots,y_n]$ corresponding to the first d minimum eigenvalues

5: Combining feature vectors into matrix $Y=[y_1,y_2,\cdots,y_n]$

First, solving the principal direction matrix U, fixing the principal component Y. Using (8) to find the partial guide [22]for U, and the derivative is equal to 0 to get:

$$\frac{\partial f}{\partial U}=\frac{\partial\left(\|X-UY\|_F^2\right)}{\partial U}=\frac{\partial((X-UY)(X-UY)^T)}{\partial U}$$
$$=-2XY^T+2U=0 \quad (9)$$

So $U=XY^T$.

Second, solving the principal component Y. Substituting into (8):

$$\min_Y \|X-XY^TY\|_F^2+\gamma tr(YHY^T)+\mu(I-YY^T) \quad (10)$$

Then finding the partial derivative for Y and letting the derivative equal to 0:

$$\frac{\partial g}{\partial Y}=\frac{\partial(tr(Y(-X^TX+\gamma H)Y^T)+\mu(I-YY^T))}{\partial Y}$$
$$=(-X^TX+\gamma H)-\mu Y=0 \quad (11)$$

Therefore, the principal component Y is the feature vector corresponding to the first d smallest eigenvalues of the matrix $(-X^TX+\gamma H)$.

B. Graph-Hessian Robust Principal Component Analysis

For the GHRPCA model, we use the Alternating direction method of multipliers (ADMM) algorithm[23] to rewrite the model to:

$$\min_{A,S,L}\|A\|_*+\lambda\|S\|_1+\gamma tr(LHL^T)\ s.t.\ X=A+S, A=L \quad (12)$$

This can decompose an entire problem into two sub-problems, which is convenient for solving. Then, using the Augmented Lagrange Multiplier method[24] to get the expression:

$$(A,S,L)=\operatorname*{argmin}_{(A,S,L)}\|A\|_*+\lambda\|S\|_1+\gamma tr(LHL^T)+\Lambda_1(X-A-S)+\frac{\beta_1}{2}\|X-A-S\|_F^2+\Lambda_2(L-A)+\frac{\beta_2}{2}\|L-A\|_F^2 \quad (13)$$

here Λ_1 and Λ_2 are the dual variables.

Then using the dual rise[25] method to solve a variable and fixing other variables. The following is a detailed process.

The first step: fixing $S, L, \Lambda_1, \Lambda_2$, solving for A.

$$A^*=\operatorname*{argmin}_A\|A\|_*+\Lambda_1(X-A-S)+\frac{\beta_1}{2}\|X-A-S\|_F^2+\Lambda_2(L-A)+\frac{\beta_2}{2}\|L-A\|_F^2$$
$$=\operatorname*{argmin}_A\|A\|_*+\frac{\beta_1}{2}\left\|X-A-S+\beta_1^{-1}\Lambda_1\right\|_F^2+\frac{\beta_2}{2}\left\|L-A+\beta_2^{-1}\Lambda_2\right\|_F^2$$
$$=\operatorname*{argmin}_A\|A\|_*+\frac{\beta_1+\beta_2}{2}\left\|L-\frac{\beta_1\left(X-S+\beta_1^{-1}\Lambda_1\right)+\beta_2\left(L+\beta_2^{-1}\Lambda_2\right)}{\beta_1+\beta_2}\right\|_F^2 \quad (14)$$

According to the definition of the singular value threshold operator of the matrix[26]:

$$\operatorname*{argmin}_X\tau\|A\|_*+\frac{1}{2}\|X-P_\Omega\|_F^2=D_\tau(P_\Omega) \quad (15)$$

among them:

$$D_\tau(X)=US_\tau(\Sigma)V^T, if\ X=U\Sigma V^T \quad (16)$$

And for the contraction operator S_τ [27], defined as:

$$S_\tau=sgn(x)max(|x|-\tau,0) \quad (17)$$

So getting the final iteration of A:

$$A^{k+1}=D_{(\beta_1+\beta_2)^{-1}}\left(\frac{\beta_1\left(X-S^K+\beta_1^{-1}{\Lambda_1}^k\right)+\beta_2\left(L^k+\beta_2^{-1}{\Lambda_2}^k\right)}{\beta_1+\beta_2}\right) \quad (18)$$

here k is the number of iterations.

The second step: fixing $A, L, \Lambda_1, \Lambda_2$, solving for S.

$$S^*=\operatorname*{argmin}_S\lambda\|S\|_1+\Lambda_1(X-A-S)+\frac{\beta_1}{2}\|X-A-S\|_F^2$$
$$=\operatorname*{argmin}_S\lambda\|S\|_1+\frac{\beta_1}{2}\left\|X-A-S+\beta_1^{-1}\Lambda_1\right\|_F^2$$
$$=S_{\frac{\lambda}{\beta_1}}\left(X-A+\beta_1^{-1}\Lambda_1\right) \quad (19)$$

Thus $S^{k+1}=S_{\frac{\lambda}{\beta_1}}\left(X-A^{k+1}+\beta_1^{-1}{\Lambda_1}^k\right)$ (20)

The third step: fixing $A, S, \Lambda_1, \Lambda_2$, solving for L.

$$L^*=\operatorname*{argmin}_L\gamma tr(LHL^T)+\Lambda_2(L-A)+\frac{\beta_2}{2}\|L-A\|_F^2$$
$$=\operatorname*{argmin}_L\gamma tr(LHL^T)+\frac{\beta_2}{2}\left\|L-A+\beta_2^{-1}\Lambda_2\right\|_F^2 \quad (21)$$

This is a smoothing function, so using the optimality condition to find a closed-form solution of L [16], and using the projection conjugate gradient method to iterate L [28][29]:

$$L^{k+1}=\beta_2(\gamma H+\beta_2 I)^{-1}\left(A^{k+1}-\frac{{\Lambda_2}^k}{\beta_2}\right) \quad (22)$$

The final step: iterating Lagrange multipliers Λ_1 and Λ_2, fixing A, L, S.

$${\Lambda_1}^{k+1}={\Lambda_1}^k+\beta_1\left(X-A^{k+1}-S^{k+1}\right) \quad (23)$$
$${\Lambda_2}^{k+1}={\Lambda_2}^k+\beta_2\left(L^{k+1}-A^{k+1}\right) \quad (24)$$

Algorithm 2 GHRPCA's ADMM algorithm

1: Input X,H,λ,γ

2: Internalize: $k=1$; $\lambda = 1/\sqrt{max(p,n)}$; $\gamma \geq 0$; $\beta_1 = 1.25/\max eigenvector(X)$; $\beta_2 = 0.01$

3: Internalize: $A = zeros(p,n)$; $S = zeros(p,n)$; $L = zeros(p,n)$

4: Internalize: $\Lambda_1 = X - A - S$; $\Lambda_2 = L - A$

5: while not converged do

6: $A^{k+1} = D_{(\beta_1+\beta_2)^{-1}}\left(\frac{\beta_1\left(X-S^K+{\beta_1}^{-1}{\Lambda_1}^k\right)+\beta_2\left(L^k+{\beta_2}^{-1}{\Lambda_2}^k\right)}{\beta_1+\beta_2}\right)$

7: $S^{k+1} = S_{\frac{\lambda}{\beta_1}}\left(X - A^{k+1} + {\beta_1}^{-1}{\Lambda_1}^k\right)$

8: $L^{k+1} = \beta_2(\gamma H + \beta_2 I)^{-1}\left(A^{k+1} - \frac{{\Lambda_2}^k}{\beta_2}\right)$

9: ${\Lambda_1}^{k+1} = {\Lambda_1}^k + \beta_1\left(X - A^{k+1} - S^{k+1}\right)$ and

${\Lambda_2}^{k+1} = {\Lambda_2}^k + \beta_2\left(L^{k+1} - A^{k+1}\right)$

10: end while

11: Output A,S

C. Calculation of the Hessian Matrix

In this section, we brief the solution steps for the Hessian matrix to get the local geometry of the sample in the following steps[17]:

Step 1: Construct a neighbor matrix $\mathcal{M}^i$. Find the k-nearest neighbors' collection $\mathcal{N}_i$. For each neighborhood $\mathcal{N}_i$ for each sample point $x_i, i = 1,2,\cdots,n$. Form a matrix $\mathcal{M}^i \in R^{k\times n}$, the j-th row of the matrix $\mathcal{M}^i$ is $x_j - \overline{x_i}, \overline{x_i} = Ave\{x_j : x_j \in \mathcal{N}_i\}$.

Step 2: Obtain tangential coordinates. Perform Singular value decomposition on $\mathcal{M}^i$ to obtain matrices U, D and V. The first d columns of U give the tangent coordinates of the sample points in $\mathcal{N}_i$.

Step 3: Calculate the Hessian estimator. Perform a least -squares estimation on Hessian get H^i. If f is a smoothing function $f: M \to R$, $f_j = (f(m_i))$,then all high-dimensional observations in $\mathcal{N}_i$ are mapped from f to f_j The set of real numbers constitutes a vector v^i. Then, all high-dimensional observation samples in $\mathcal{N}_i$ consist of a real set of f to f_j to form a vector v^i. Then, the product $H^i v^i$ gives a vector of length $d(d+1)/2$, and the element $\partial^2 f/\partial u_i \partial u_j$ in the vector is an estimate of the Hessian matrix.

Step 4: Calculate the quadratic form $\mathcal{H}$. According to $\mathcal{H}_{ij} = \sum_l \sum_r \left(\left(H^l\right)_{r,i}\left(H^l\right)_{r,j}\right)$ to calculate the symmetric matrix $\mathcal{H}_{ij}$. The matrix $H^l \in R^{d(d+1)/2\times k}$ also refers to the Hessian estimate of the neighbor matrix. The row r represents the corresponding element in the Hessian matrix, and the column i represents the corresponding neighbor points.

D. Computational Complexity Analysis

In the process of solving the Graph-Hessian Principal Component Analysis, the main cost is the singular value decomposition to obtain the eigenvalues and eigenvectors. Therefore, the complexity of this model is similar to that of PCA.

TABLE I. DATASETS INFORMATION COMPARISON

Dataset	Class	Dimensionality	Size
USPS	10	256	9298
YALE	15	1024	165
COIL20	20	16384	1440

The Graph-Hessian Robust Principal Component Analysis is solved by the Augmented Lagrange Multiplier methods. This algorithm is sufficient for good accuracy in our model with a small number of iterations. The complexity of nuclear norm proximal computation for updating A is $\mathcal{O}(np^2 + p^3)$ and the computational complexity of the Conjugate Gradient method for updating L is $\mathcal{O}(np)$.Thus, the main cost of each iteration corresponds to the computation of nuclear proximal operator.

V. EXPERIMENTS

In this paper, we use the model to perform an experiment based on principal component analysis: K-means clustering based on lossless and lossy data in low-dimensional space. Our experiment is to show the robustness of our proposed model to noise and the generalization ability for different types of datasets. Our experiments were conducted on three different types of datasets, and we thought that four different levels of data corruption were introduced and compared with the five existing dimensionality reduction models.

A. Datasets Introduction

This experiment involves three databases: the handwritten digital dataset USPS[30], the face dataset YALE[31], and the object image dataset COIL20[32]. The data sets used are all relatively high-use datasets, and the tag information for each dataset is known.

Handwritten Digital Dataset USPS[30]: Fully known as the United States Postal Service handwritten digital dataset, consisting of 10 handwritten digital scan crops from 0 to 9, comprising a total of 9298 8-bit grayscale images, the image background is black, the size of each image is 16×16 pixels. Part of the sample pictures are shown in Fig. 2.

Face Dataset YALE[31]: Created by Yale Center for Computing Vision and Control. This dataset contains 165 images in GIF format. A total of 15 people participated in the film. Pictures of 10 people under certain conditions are selected, as shown in Fig. 3. Each person has 11 different facial expressions or configured images, each with different facial expressions. Or configure one: center light, with glasses, happy, left light, no glasses, normal, right light, sad, sleepy, surprised and wink, each picture is 32×32 size.

Object image dataset COIL20[32]: Fully called Columbia University Image Library dataset, this dataset contains 1,440 sheets, shooting 20 objects from different angles, shooting an image every 5 degrees, 72 images per object. Each image is uniformly sized, with a black background and a size of 128×128 pixels. A partial type of pictures are shown in Fig. 4.

B. Parameter Selection

For the GHPCA model, there is one parameter: Graph regularization coefficient γ. The value of the parameter γ is determined by means of cross-validation.

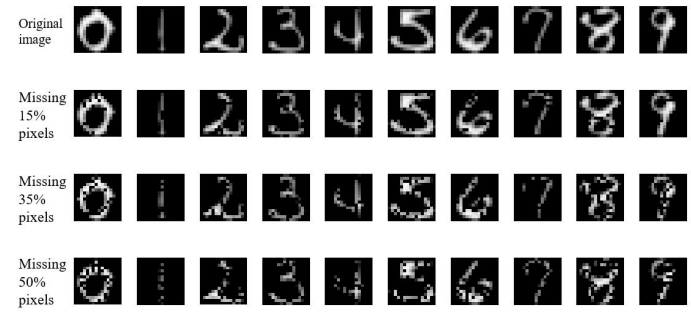

Fig. 2. Sample images from the USPS dataset.

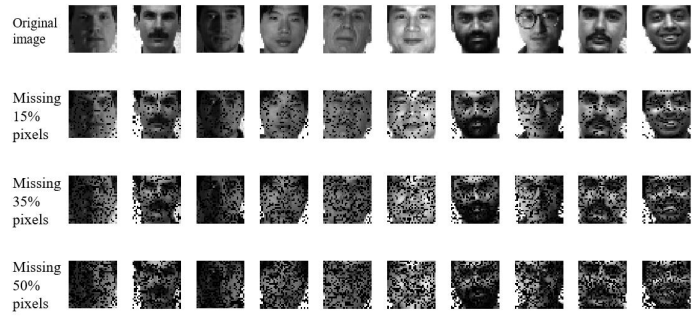

Fig. 3. Sample images from the YALE dataset.

Fig. 4. Sample images from the COIL20 dataset.

For the GHRPCA model, there are two parameters: Sparsity λ and Graph regularization coefficient γ. The parameter λ can be set approximately equal to $1/\sqrt{max(p,n)}$[5], where n is the number of data samples and p is the dimensionality of the sample. After being fixed, the value of the parameter γ is determined by cross-validation.

C. K-means Clustering Experiment

In this paper, the K-means clustering experiment based on PCA[33]is used to compare the robustness of different models to lossy information and the generalization ability for different datasets. We use all the data from each dataset for unsupervised clustering experiment. The label information of each sample is known and the total number of categories is also known.

a) Data Processing

In each dataset, we introduce four different degrees of missing: no missing, randomly missing 15% of the pixel value, randomly missing 35% of the pixel value and randomly missing 50% of the pixel value. Generating a uniformly distributed pseudo-random integer as the position of the pixel, and writing 0 to the original data matrix.

TABLE II. THE CLUSTERING CORRECT RATE (%) WITH DIFFERENT DIMENSIONALITY REDUCTION MODELS

Data Set	Model / Miss	*K-means*	*PCA*	*GL PCA*	*GH PCA*	*RPCA*	*RPCAG*	*GH RPCA*
USPS	*0%*	59.43	61.05	65.42	**66.15**	71.79	79.77	**80.29**
	15%	66.77	68.31	72.01	**73.02**	71.63	78.29	**79.32**
	35%	67.72	68.81	71.49	**73.01**	69.90	77.23	**77.78**
	50%	68.21	70.79	72.50	**73.78**	71.18	71.62	**73.40**
YALE	*0%*	46.79	44.00	48.48	**51.27**	50.91	54.79	**55.52**
	15%	40.85	39.15	41.09	**45.09**	52.73	54.55	**56.48**
	35%	34.79	36.36	37.33	**38.79**	44.61	46.18	**47.52**
	50%	30.30	32.97	34.55	**36.61**	38.42	40.73	**41.82**
COIL20	*0%*	59.89	61.08	61.15	**62.22**	62.13	65.65	**65.96**
	15%	59.44	61.07	62.01	**62.85**	62.29	64.93	**66.08**
	35%	58.06	59.51	60.01	**61.85**	64.96	65.06	**66.19**
	50%	55.22	59.49	59.74	**60.86**	61.43	62.56	**63.04**

For each dimensionality reduction model, after the introduction of the missing, we perform preprocessing of zero mean and unit standardization based on sample features for all three datasets.

b) Repeatability Experiment Setup

For each degree of missing, we performed 5 different randomly independent replicates experiments (the no missing experiment was also performed 5 times). After the dimensionality reduction is completed, each group of data is subjected to 100 times K-means clustering experiments to calculate the correct rate of the experiment.

c) Clustering Correct Rate

The clustering correct rate is calculated by comparing the original label with the label information obtained after clustering. The result of each missing was taken as the minimum of the 100 times experiments, and the final result of the experiment was taken as the average of 5 randomly missing experiments.

D. Experimental Result

K-means experiments were performed on three different databases for different dimensionality reduction models. All the specific experimental data are shown in Table Ⅱ. The data from

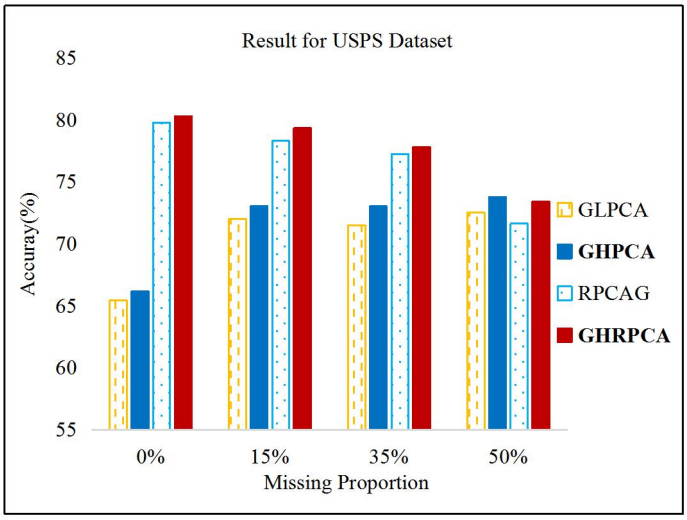

Fig. 5. The clustering correct rate (%) with Hessian-based models and Laplacian-based models on USPS dataset.

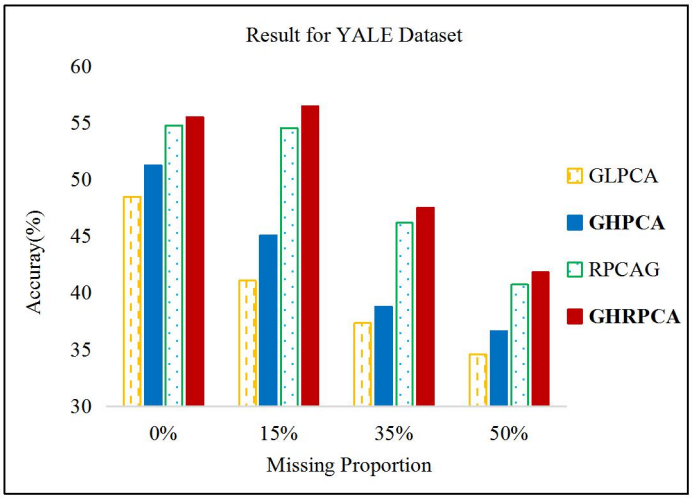

Fig. 6. The clustering correct rate (%) with Hessian-based models and Laplacian-based models on YALE dataset.

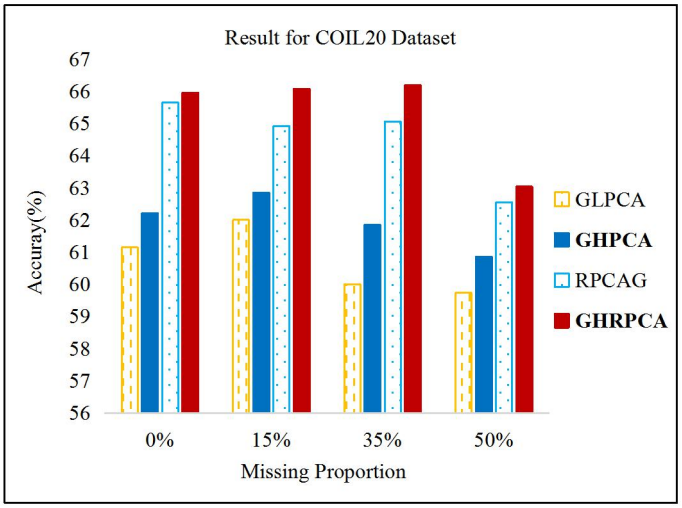

Fig. 7. The clustering correct rate (%) with Hessian-based models and Laplacian-based models on COIL20 dataset.

our model has been boldly displayed. We can see that our Hessian models have a higher accuracy rate than other models. A detailed comparison of the Hessian-based model of the same category with the Laplacian-based model is shown in Fig. 5, Fig. 6, Fig. 7.

Fig. 5 shows a bar chart showing the correct clustering rate of our Hessian-based models and Laplacian-based models on the USPS dataset. Fig. 6 shows a bar chart of the results of our Hessian-based models and Laplacian-based models on the YALE dataset. Fig. 7 shows a bar chart of the results on the COIL20 dataset. The X-axis indicates the proportion of data missing: 0%, 15%, 35%, 50%, and the Y-axis represents the correct rate of the experimental results. Models involved in each figure: GLPCA, GHPCA, RPCAG and GHRPCA. The model represented by each column is shown on the right side of the figure, and our two models are represented by solid columns.

We can see that, whether in the handwritten digital, face or object image dataset, our proposed GHPCA has higher accuracy than GLPCA under four different proportions of data missing, our proposed GHRPCA has higher clustering accuracy than RPCAG. We can get our model to have better ability for different features, such as face features, digital features and object features. In other words, Hessian regularization is also suitable for principal component analysis models.

And we prove that Hessian can get more useful information than Laplacian. Therefore, we can gain that our models can still obtain more abundant local structure information after dimensionality reduction in different data, even in the case of a great deal of data missing. We have used this ability of Hessian-based models to get a better clustering effect.

Thence we conclude that our models are superior than other models on different datasets, which shows that they can keep more local geometric relations in low-dimensional space, more robust to noise and have better generalization ability.

VI. Conclusion

Extracting useful information from destroyed high-dimensional data and reduce the data dimensionality has become one of the main tasks of machine learning. Principal Component Analysis is the simplest and most popular linear dimensionality reduction method. In this paper, we use the Hessian to construct the spectral regularization into the framework of Principal Component Analysis and propose two models: Graph-Hessian Principal Component Analysis (GHPCA) and Graph-Hessian Robust Principal Component Analysis (GHRPCA).

Since the null space of the Hessian is richer than the Laplacian, it is more abundant and accurate in terms of structural information between the reaction samples. By the K-means clustering experiments on USPS handwritten digital dataset, YALE face dataset and COIL20 object image dataset, by comparing with several methods of PCA, GLPCA, RPCA and RPCAG, it proves that our proposed Hessian-based Principal Component Analysis models are superior to other Principal Component Analysis models. Our models can get more abundant information. In future work, we will further

study its characteristics in order to increase accuracy and reduce the time cost of operation.

ACKNOWLEDGMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61671480, in part by the Fundamental Research Funds for the Central Universities, China University of Petroleum (East China) under Grant 18CX07011A and YCX2019080, in part by the Science and Technology Development Fund, Macau SAR (File no. 189/2017/A3), and by the Research Committee at University of Macau under Grants MYRG2016-00123-FST and MYRG2018-00136-FST.

REFERENCES

[1] C. P. Chen and C. Y. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information sciences*, vol. 275, pp. 314-347, 2014.

[2] V. Rene, Y. Ma and S. Sastry, *Generalized principal component analysis*. Springer, 2016.

[3] I. Jolliffe, *Principal component analysis*. Springer, 2011.

[4] C. Ding, D. Zhou, X. He and H. Zha, "R1-PCA: rotational invariant L1-norm principal component analysis for robust subspace factorization," *Proceedings of the 23rd International Conference on Machine Learning*. *ACM*, vol. 281-288, 2006.

[5] E. J. Candès, X. Li, Y. Ma and J. Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, p. 11, 2011.

[6] H. Zou, T. Hastie and R. Tibshirani, "Sparse principal component analysis," *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265-286, 2006.

[7] A. K. Seghouane and A. Iqbal, "The adaptive block sparse PCA and its application to multi-subject FMRI data analysis using sparse mCCA," *Signal Processing*, vol. 153, pp. 311-320, 2018.

[8] H. Shen and J. Z. Huang, "Sparse principal component analysis via regularized low rank matrix approximation," *Journal of Multivariate Analysis*, vol. 99, no. 6, pp. 1015-1034, 2008.

[9] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611-622, 1999.

[10] N. Lawrence, "Probabilistic non-linear principal component analysis with Gaussian process latent variable models," *Journal of Machine Learning research*, vol. 6, no. Nov, pp. 1783-1816, 2005.

[11] B. Schölkopf, A. Smola and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, no. 5, pp. 1299-1319, 1998.

[12] M. Ding, Z. Tian and H. Xu, "Adaptive kernel principal component analysis," *Signal Processing*, vol. 90, no. 5, pp. 1542-1553, 2010.

[13] X. Du, Y. Yan, P. Pan, G. Long and L. Zhao, "Multiple graph unsupervised feature selection," *Signal Processing*, vol. 120, pp. 754-760, 2016.

[14] Z. Zhang and K. Zhao, "Low-rank matrix approximation with manifold regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1717-1729, 2013.

[15] B. Jiang, C. Ding and J. Tang, "Graph-Laplacian PCA: Closed-form solution and robustness," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3492-3498, 2013.

[16] N. Shahid, V. Kalofolias, X. Bresson, M. Bronstein and P. Vandergheynst, "Robust principal component analysis on graphs," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2812-2820, 2015.

[17] D. L. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proceedings of the National Academy of Sciences*, vol. 100, no. 10, pp. 5591-5596, 2003.

[18] K. I. Kim, F. Steinke and M. Hein, "Semi-supervised regression using Hessian energy with an application to semi-supervised dimensionality reduction," *Advances in Neural Information Processing Systems*, pp. 979-987, 2009.

[19] W. Liu and D. Tao, "Multiview hessian regularization for image annotation," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2676-2687, 2013.

[20] W. Liu, H. Liu, D. Tao, Y. Wang and K. Lu, "Multiview Hessian regularized logistic regression for action recognition," *Signal Processing*, vol. 110, pp. 101-107, 2015.

[21] J. Eells and L. Lemaire, *Selected topics in harmonic maps*. *American Mathematical Soc*, 1983.

[22] K. B. Petersen and M. S. Pedersen, *The matrix cookbook*. *Technical University of Denmark*, vol. 7, no. 15, p. 510, 2008.

[23] S. Boyd, N. Parikh, E. Chu, B. Peleato and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1-122, 2011.

[24] Z. Lin, M. Chen and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:*1009.5055, 2010.

[25] X. Yuan and J. Yang, "Sparse and low-rank matrix decomposition via alternating direction methods," *preprint*, vol. 12, no. 2, 2009.

[26] J. F. Cai, E. J. Candès and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956-1982, 2010.

[27] S. J. Wright, R. D. Nowak and M. A. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479-2493, 2009.

[28] P. H. Calamai and J. J. Moré, "Projected gradient methods for linearly constrained problems," *Mathematical Programming*, vol. 39, no. 1, pp. 93-116, 1987.

[29] Z. Liang, Y. Li and T. Zhao, "Projected gradient method for kernel discriminant nonnegative matrix factorization and the applications," *Signal Processing*, vol. 90, no. 7, pp. 2150-2163, 2010.

[30] J. J. Hull, "A database for handwritten text recognition research," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 5, pp. 550-554, 1994.

[31] R. Gross, "Face databases," *Handbook of face recognition*. Springer, New York, NY, pp. 301-327, 2005.

[32] S. A. Nene, S. K. Nayar and H. Murase, "Columbia object image library (coil-20)," 1996.

[33] C. Ding and X. He, "K-means clustering via principal component analysis," *Proceedings of the 21rd International Conference on Machine Learning*. *ACM*, vol. 29, 2004.