21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Orthogonal Isometric Projection

Yali Zheng[1], Yuan Yan Tang[1,2], Bin Fang[1], Taiping Zhang[1]
1. Chongqing University, Chongqing, China 2. University of Macau, Macau, China
zhengyl, fb, tpzhang@cqu.edu.cn, yytang@umac.mo

Abstract

In contrast with Isomap, which learns the low-dimension embedding, and solves problem under the classic Multi-dimension Scaling (MDS) framework, we propose a dimensionality reduction technique, called Orthogonal Isometric Projection (OIP), in this paper. We consider an explicit orthogonal linear projection by capturing the geodesic distance, which is able to handle new data straightforward, and leads to a standard eigenvalue problem. And we extend our method to Sparse Orthogonal Isometric Projection (SOIP), which can be solved efficiently using LARS. Numerical experiments are reported to demonstrate the performance of OIP by comparing with a few competing methods.

1. Introduction

Researchers are interested in reducing dimensions of digital data is because of the existence of digital information redundancy. It is well known that extracting efficient features from data can improve object classification and recognition, and simplify the visualization of data. Manifold learning theory [4, 3, 11, 12] was introduced into dimensionality reduction field in early 20 century, which assumed that a low-dimension manifold is embedded in high-dimension data. Researchers pay lots of attention to discovery, and prove that a low-dimension manifold exists. Isomap was proposed by Tenenbaum et al. [4], in which geodesic distance was used to capture the global structure in high dimension, and can be solved under MDS framework. Roweis and Saul [3] proposed a nonlinear dimensionality reduction method, Locally Linear Embedding (LLE), which aimed at preserving the same local configuration of each neighborhood in low dimensional space as in high dimensional space. He and Niyogi [2] found an optimal linear approximations to eigenfunctions of the nonlinear Laplacian Eigenmap, called Local Preserving Projection (LPP), and gave the justification in the paper with the Graph Laplacian theory. Cai et al. [5] proposed a variation of Laplacianface (LPP) — Orthogonal Laplacianface (OLPP), which iteratively computed the orthogonal eigenvectors to compose the projection matrix. Kokiopoulou and Saad [6] analyzed and compared LPP and OLPP, and proposed an Orthogonal Neighborhood Preserving Projections (ONPP). It can be thought of as an orthogonal version of LLE, but projections are learned explicitly as a standard eigenvalue problem.

Dimensionality reduction techniques is either to seek for a representation of data in low dimension to benefit the data analysis or to map data from high-dimensional space to low-dimensional space through an explicit linear or nonlinear projection learned from a training dataset. Multidimensional Scaling (MDS) [1] is a classic data embedding technique, and considers preserving the pairwise distance to obtain the low dimension configuration. Principal component analysis (PCA) can be used as a projection method, which learns a linear projection by maximizing the variance of data in low dimension. PCA is identical to Classical Multidimensional Scaling if euclidean distance is used [1], but PCA learns the projection. Linear Discriminant Analysis (LDA) maximizes the ratio of between-class variance to the within-class variance to determine an explicit projection as well.

In this paper, we are motivated by Isomap [4] and Isometric projection [8], and propose a linear projection method, called Orthogonal Isometric Projection, which is a variation of Isometric Projection. Cai proposed Isometric Projection [8] addressed the same purpose as ours. However, in this paper, we constrain the projection is orthogonal which differs from Cai's, and solve a standard eigenvalue problem. The main difference between our method and Cai's orthogonal version of Isometric Projection is that: we build a reasonable objective function, and solve the optimization in a standard eigenvalue problem, while Cai solved the problem in a generalized eigenvalue problem. We extend our method to the sparse orthogonal isometric projection. We test our method on USPS data set. In the following

978-4-9906441-1-6

section we will briefly introduce Cai's Isometric Projection to demonstrate the advantages of our algorithm.

2. A brief review of Isomap and Isometric Projection

2.1 Isomap

Isomap was proposed by Tenenbaum and et al. [4], and is one of most popular manifold learning techniques. It aims to obtain an euclidean embedding of points such that the geodesic distance in high dimensional space gets close to the euclidean distance between each pair of points. The mathematical formulation is

$$\min \sum_{i,j} (d_G(x_i, x_j) - d_E(y_i, y_j))^2 \tag{1}$$

d_G is the geodesic distance, which is defined locally to be the shortest path on the manifold, d_E is the euclidean distance, and in a matrix form:

$$\min \|\tau(D_G) - \tau(D_E)\|_2 \tag{2}$$

D_G is the geodesic distance matrix, D_E is the euclidean distance matrix, τ is an operation which converts the euclidean distance into an inner product form. The problem is solved under the MDS framework. Isomap makes an assumption that a manifold existing in high dimension space, and applies the geodesic distance to measure the similarity of each point pair. However, if insufficient samples are given or the data is noised, the intrinsic geometry of the data is difficult to be captured by constructing the neighborhood graph.

2.2 Isometric Projection

Cai et al. [8] extended Isomap algorithm to learn a linear projection by solving a spectral graph optimization problem. Suppose that $Y = V^T X$, they minimized the objective function,

$$\min \|\tau(D_G) - X^T V V^T X\|_2 \tag{3}$$

To make the problem tractable, they imposed a constraint $V^T X X^T V = I$, and rewrote the minimization problem as

$$\begin{array}{ll} \arg\max\limits_{V} & tr(V^T X \tau(D_G) X^T V) \\ \text{s.t.} & V^T X X^T V = I \end{array}$$

which is equivalent to a generalized eigenvalue problem

$$X\tau(D_G)X^T V = \lambda X X^T V$$

To solve the problem efficiently in computation cost, Cai also applied the regression in [8, 13] over Y and X called spectral regression (SR). Y is computed first, which is the eigenvector of $\tau(D_G)$, then

$$\mathbf{a} = \arg\min_{\mathbf{a}} \sum_{i=1}^{m} (\mathbf{a}^T x_i - y_i)^2 + \alpha \|\mathbf{a}\|^2 \tag{4}$$

The condition $V^T X X^T V = I$ constrained that low-dimension embedding of points is orthogonal, namely, $Y^T Y = I$.

3. Orthogonal Isometric Projection

The main idea of orthogonal isometric projection is to seek an orthogonal mapping over the training data set so as to best preserve the geodesic distance on a neighborhood graph, and learn a linear projection under the general Isomap framework, but has a different and reasonable constraint that projections are orthogonal.

3.1 The Objective Function of OIP

Under the Isomap framework, it minimizes the objective function in Equation 2. In math $\tau(D_G) = -CWC/2$, and C is the centering matrix defined by $C = I_n - 1/N \cdot e_N e_N^T$, where $e_N = [1, ..., 1]_N^T$, W is a Dijkstra distance matrix based on K nearest neighbor graph over all data points. Let $f(V) = \|\tau(D_G) - X^T V V^T X\|_2$, we are seeking for a linear projection:

$$\min_{V} f(V) \tag{5}$$

Let $S = \tau(D_G)$, which is a known neighborhood graph constructed from the given data set. We have

$$\begin{aligned} f(V) &= tr((S - X^T V V^T X)^T (S - X^T V V^T X)) \\ &= tr(S^T S) + tr((X^T V V^T X)^T (X^T V V^T X) \\ &\quad -2S^T (X^T V V^T X)) \\ &= tr((X^T V V^T X)^T - 2S^T)(X^T V V^T X) \\ &\quad +tr(S^T S) \end{aligned}$$

So the objective function equation 5 is equivalent to

$$\begin{array}{ll} \min & tr(V^T X((X^T V V^T X)^T - 2S^T) X^T V) \\ \text{s.t.} & V^T V = I \end{array}$$

Let $M = X(X^T X - 2S^T)X^T$, then the problem becomes to

$$\begin{array}{ll} \min & tr(V^T M V) \\ \text{s.t.} & V^T V = I \end{array}$$

Algorithm 1 Orthogonal Isometric Projection

1: Construct a neighborhood graph G over the data points. Compute the Dijkstra distances matrix $W(i,j) = d_G(i,j)$ over the graph between every point pair. $d_G(i,j)$ is the shortest path of i and j, otherwise $d_G(i,j) = \inf$.
2: Compute $\tau(D_G) = -CWC/2$, C is the centering matrix, and $C = I_n - 1/N \cdot e_N e_N^T$.
3: Compute the eigenvectors of $M = X(X^TX - 2\tau(D_G)^T)X^T$. V is determined by eigenvectors of M associated with p smallest eigenvalues.

which leads to a standard eigenvalue problem,

$$MV = \lambda V \tag{6}$$

V is determined by eigenvectors of M corresponding to p smallest eigenvalues.

3.2 The Algorithm of OIP

First we need to construct the neighborhood graph G. As all we know, there are two options:

- if $x_j \in \mathbb{N}(x_i)$, $\mathbb{N}(x_i)$ is the k nearest neighbor of x_i,
- if $\|x_i - x_j\|^2 \leq \epsilon$, ϵ is a small number as a threshold.

the edge between i and j weighted by gaussian kernel is defined as $e^{-\|x_i - x_j\|^2}$.

We summarize the algorithm in Alg 1.

4 Sparse Orthogonal Isometric Projection

Our problem essentially solves a standard optimization in Equation 6, which also can be incorporated into a regression framework in the way of Sparse PCA [10]. The optimal V is the eigenvectors with respect to the maximum eigenvalues of Equation 6. Since M is a real symmetric matrix, so M can be decomposed into $\tilde{X}\tilde{X}^T$. Suppose the rank of X is r, and SVD($\tilde{X}$) is

$$\tilde{X} = \tilde{U}\tilde{\Sigma}\tilde{V}^T,$$

it is easy to verify that the column vectors in $\tilde{U}$ are the eigenvectors of $\tilde{X}\tilde{X}^T$. Let $Y = [y_1, y_2, \cdots, y_r]_{n\times r}$, each row vector is a sample vector in r-dimensional subspace, and $V = [v_1, v_2, \cdots, v_r]_{m\times r}$. Therefore, the projective functions of OIP are solved by the linear systems:

$$X^T v_p = y_p, p = 1, 2, \cdots, r.$$

v_p is the solution of the regression system:

$$v_p = \arg\min_v \sum_{i=1}^{n} (v^T x^i - y_p^i)^2,$$

where y_p^i is the ith element of y_p. Similar to Zou *et al.* [10], we can get the sparse solutions by adding L_1 regularization:

$$v_p = \arg\min_v \sum_{i=1}^{n} (v^T x^i - y_p^i)^2 + \alpha \sum_{j=1}^{m} \|v_p^j\|$$

where v_p^j is the jth element of v_p. The regression can be solved efficiently using LARS algorithm.

5. Experiments

USPS is a well known handwritten digits corpus from US postal service. It contains normalized gray scale images of size 16×16, and totally 9298 samples with 256 features. Fig. 1 shows some examples of USPS dataset. We evaluate our algorithm on USPS, which are downloaded from public website[1], and compare with PCA, LDA, LPP, IP, and IP+SR, and demonstrate the average accuracy and average error rates in the section. A human error rate estimated to be 2.37% shows that it is a hard task over USPS dataset [9]. We randomly sample 25 times from the datasets as the training sets with varying rates from 20% to 80%, and the rest are used for testing. We map points in test sets by projections learned from training sets, and apply the nearest neighbor to determine categories labels. We also demonstrate the dimensions versus average error rate by half training and half testing. Assume that l_i is the ground truth, b_i is the label assigned after dimensionality reduction by methods, N_{test} is the number of test samples,

$$Acc = \frac{1}{N} \sum_{j=1}^{N} \frac{\sum_{i=1}^{N_{test}} \delta(l_i, proj(b_i))}{N_{test}},$$

and $N = 25$ in our experiment.

We compare our method with IP, IP+SR, LDA, and LPP in Table 1. Our method outperforms all other methods. The bolds are from our method, $\pm$ shows the square root of the variance. With the number of training samples, the classification precision increases as what we expect. Fig. 2 shows with the number of dimensions increases, the average classification error decreases.

[1] http://www.zjucadcg.cn/dengcai/Data/TextData.html

Table 1. Comparison on USPS

Train ratio	LDA	LPP	IP	IP+SR	OIP
20%	89.13 ± 0.34	92.95 ± 0.36	92.11 ± 0.39	93.90 ± 0.50	**95.10** ± 0.20
30%	90.39 ± 0.32	94.47 ± 0.30	93.61 ± 0.33	94.96 ± 0.36	**95.93** ± 0.17
40%	91.18 ± 0.36	95.20 ± 0.27	94.48 ± 0.28	95.69 ± 0.36	**96.40** ± 0.20
50%	91.54 ± 0.32	95.75 ± 0.29	94.85 ± 0.42	95.91 ± 0.34	**96.65** ± 0.26
60%	91.97 ± 0.32	96.14 ± 0.24	95.21 ± 0.31	96.14 ± 0.35	**97.01** ± 0.25
70%	92.02 ± 0.49	96.43 ± 0.33	95.61 ± 0.35	96.59 ± 0.32	**97.17** ± 0.23
80%	92.19 ± 0.58	96.71 ± 0.42	95.97 ± 0.40	96.74 ± 0.40	**97.35** ± 0.34

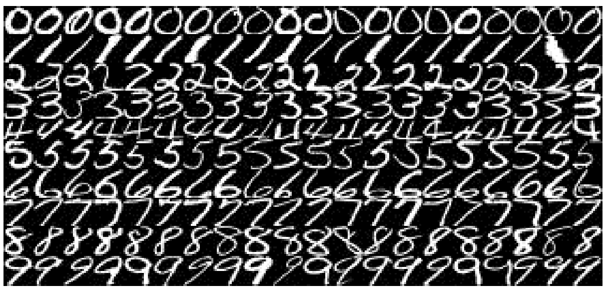

Figure 1. USPS examples

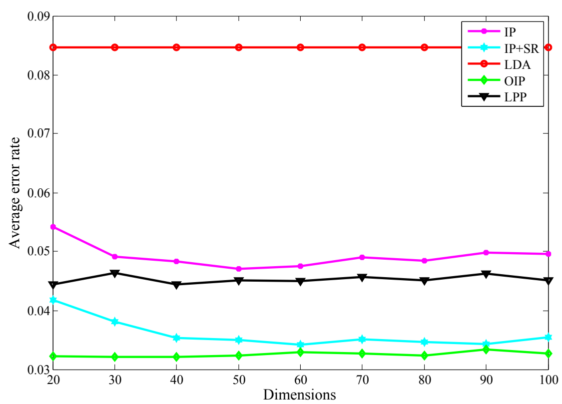

Figure 2. Dimensions vs. average classification error on USPS dataset.

References

[1] W. S. Torgerson, Multidimensional scaling: I. Theory and method, Psychometrika, **17(4)**(1952), pp. 401–419.

[2] X. He and P. Niyogi, Locality preserving projections, Proceeding of Advances in Neural Information Processing Systems, 2003, pp. 153–160.

[3] S. T. Roweis and L. K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science, **290**(2000), pp. 2323–2326.

[4] J. B. Tenenbaum, V. Silva and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science, **290**(2000), pp. 2319–2323.

[5] D. Cai, X. He, J. Han and H.J. Zhang, Orthogonal laplacianfaces for face recognition, IEEE Transactions on Image Processing, **15(11)**(2006), pp. 3608–3614.

[6] E. Kokiopoulou and Y. Saad, Orthogonal neighborhood preserving projections: a projection-based dimensionality reduction technique, IEEE transactions on Pattern Analysis and Machine Intelligence, **29(12)**(2007), pp. 2143–2156.

[7] X. He, S. Yan, Y. Hu, P. Niyogi and H.J. Zhang, IEEE Transactions on Pattern Analysis and Machine Intelligence, **27(3)**(2005), pp. 328–340.

[8] D. Cai, X. He and J. Han, Isometric projection, Proceedings of the National Conference on Artificial Intelligence, 2007, pp. 528–533.

[9] I. Chaaban and M. Scheessele, Human performance on the USPS database, 2007, Technique report, Indiana University South Bend.

[10] H. Zou, T. Hastie and R. Tibshirani, Sparse principal component analysis, Journal of computational and graphical statistics, **15(2)**(2006), pp. 265–286.

[11] J.A. Lee and M. Verleysen, Nonlinear dimensionality reduction of data manifolds with essential loops, Neurocomputing, **67**(2005), pp. 29–53.

[12] H. Choi and S. Choi, Robust kernel isomap, Pattern Recognition, **40(3)**(2007), pp. 853–862.

[13] D. Cai, X. He and J. Han, SRDA: An efficient algorithm for large-scale discriminant analysis, IEEE Transactions on Knowledge and Data Engineering, **20(1)**(2008), pp. 1–12.