Locality Constrained Dictionary Learning for Nonlinear Dimensionality Reduction

Yin (Joe) Zhou and Kenneth Barner

UNIVERSITY OF DELAWARE

IBM T. J. Watson Seminar
Outline

1. Introduction
   - Overview
   - NLDR and Examples
   - Motivations
   - Proposed Framework

2. Background Knowledge

3. Proposed Approach
   - Problem Formulation
   - Locality-Constrained Dictionary Learning

4. Experiments
   - Synthetic Datasets
   - Real-world Datasets

5. Conclusion and Future Work

6. Appendix
Outline

1 Introduction
   - Overview
   - NLDR and Examples
   - Motivations
   - Proposed Framework

2 Background Knowledge

3 Proposed Approach
   - Problem Formulation
   - Locality-Constrained Dictionary Learning

4 Experiments
   - Synthetic Datasets
   - Real-world Datasets

5 Conclusion and Future Work

6 Appendix
Overview

- Efficiently processing large-scale high-dimensional data is a challenging problem in machine learning.
- In this work, we propose a framework to accelerate nonlinear dimensionality reduction (NLDR) efficiency.

Experiments show that our method can improve the dimensionality reduction efficiency by more than 2 orders of magnitude.
Nonlinear Dimensionality Reduction

Nonlinear Dimensionality Reduction (NLDR) $\iff$ Manifold Learning

NLDR estimates the intrinsic low-dimensional manifold $\mathcal{N}$

Representative NLDR algorithms are: ISOMAP, Locally Linear Embedding, Laplacian Eigenmap, Local Tangent Space Alignment, etc.
Pose and Illumination Direction Estimation

Figure: 2D manifold obtained by ISOMAP over 698 face images of dimension 4096.

- Improve face recognition accuracy
- Facilitate human-computer interaction
Medical Imaging

Figure: Brain manifold. Image source: www.na-mic.org/Wiki.

- Making it easier in searching and browsing large database
- An effective tool in clinical diagnosis
Motivations

Nowadays, NLDR is facing large-scale problems. For example,
Motivations

Nowadays, NLDR is facing large-scale problems. For example,

- In computer vision, image databases grow rapidly in size

![Caltech 256](a) Caltech 256 ![MIT SUN](b) MIT SUN

**Figure:** Caltech 256 Database (Griffin et al.) contains 30,607 images; MIT SUN Database (Xiao et al.) now includes 131,072 images and is of growing size.
Motivations

Nowadays, NLDR is facing large-scale problems. For example,

- In computer vision, image databases grow rapidly in size

![Figure: Caltech 256 Database (Griffin et al.) contains 30,607 images; MIT SUN Database (Xiao et al.) now includes 131,072 images and is of growing size.]

However, applying NLDR over large-scale databases causes exorbitant computational and memory complexity.
Motivations

Nowadays, NLDR is facing large-scale problems. For example,

- In computer vision, image databases grow rapidly in size

![Caltech 256](image1.png) ![MIT SUN](image2.png)

(a) Caltech 256  (b) MIT SUN

**Figure:** Caltech 256 Database (Griffin et al.) contains 30,607 images; MIT SUN Database (Xiao et al.) now includes 131,072 images and is of growing size.

However, applying NLDR over large-scale databases causes exorbitant computational and memory complexity.

- Generally, NLDR has two steps, *i.e.*, Nearest-Neighbor Graph Construction and Partial Eigenvalue Decomposition.
- Current NLDR algorithms have \( O(N^2) \) or \( O(N^3) \) computational complexity in the number of data \( N \), and
- \( O(N^2) \) memory complexity in the number of data.
To efficiently process large-scale databases, we propose the following framework for NLDR via learning a dictionary of landmark points.
# Outline

1. Introduction
   - Overview
   - NLDR and Examples
   - Motivations
   - Proposed Framework

2. Background Knowledge

3. Proposed Approach
   - Problem Formulation
   - Locality-Constrained Dictionary Learning

4. Experiments
   - Synthetic Datasets
   - Real-world Datasets

5. Conclusion and Future Work

6. Appendix
A manifold $\mathcal{M}$ of dimension $n$, or $n$-manifold is a topological space with the following properties:

1. $\mathcal{M}$ is Hausdorff
2. $\mathcal{M}$ is locally Euclidean of dimension $n$, and
3. $\mathcal{M}$ has a countable basis of open sets.
Basic Concepts about Manifold

- Manifold $\mathcal{M} \in \mathbb{R}^m$ is our data manifold in observation space.
- Manifold $\mathcal{N}$ is unobservable and can only be estimated.
- Manifold $\mathcal{M}$ is the image of intrinsic low-dimensional manifold $\mathcal{N}$ under mapping $f: \mathcal{N} \rightarrow \mathbb{R}^m$, where $n \ll m$.
- **Local geometry is preserved after mapping $f$ or $g$.**
Outline

1 Introduction
   - Overview
   - NLDR and Examples
   - Motivations
   - Proposed Framework

2 Background Knowledge

3 Proposed Approach
   - Problem Formulation
   - Locality-Constrained Dictionary Learning

4 Experiments
   - Synthetic Datasets
   - Real-world Datasets

5 Conclusion and Future Work

6 Appendix
Problem Formulation

Let \( Y = \{y_i\}_{i=1}^{N} \) be an observation set in \( \mathbb{R}^m \). Suppose all \( y_i \) reside on a smooth \( M \subset \mathbb{R}^m \), which is the image of a smooth \( n \)-manifold \( N \) under \( f : N \rightarrow \mathbb{R}^m \).

Goal: learn a codebook \( D = [d_1, \ldots, d_K] \) of \( K \) landmarks on \( M \), such that

\[
\|g(y_i) - g(D)x_i\|_2 \text{ is minimized for all } i = 1, \ldots, N
\]

where \( g(D) = [g(d_1), \ldots, g(d_K)] \), \( (K \ll N) \), \( x_i \) is a local reconstruction code for representing \( g(y_i) \).
Difficulties

In practice, however, it is infeasible to recover $g$. The reasons are:

1. The myriad of observed data causes intractable computational complexity and memory consumption.
2. The intrinsic manifold $\mathcal{N}$ is typically unknown.
Difficulties

- In practice, however, it is infeasible to recover $g$. The reasons are
  1. the myriad of observed data causes intractable computational complexity and memory consumption
  2. the intrinsic manifold $\mathcal{N}$ is typically unknown

- Without knowing $g$ explicitly, minimizing $\|g(y_i) - g(D)x_i\|_2$ for all $i$ on $\mathcal{N}$ becomes impractical.
Difficulties

- In practice, however, it is infeasible to recover \( g \). The reasons are
  1. the myriad of observed data causes intractable computational complexity and memory consumption
  2. the intrinsic manifold \( \mathcal{N} \) is typically unknown

- Without knowing \( g \) explicitly, minimizing \( \|g(y_i) - g(D)x_i\|_2 \) for all \( i \) on \( \mathcal{N} \) becomes impractical.

- Therefore, we need to establish a relationship between the approximation problem among latent variables (\( i.e., g(y_i) \) and \( g(D) \)) and the approximation problem among observation variables (\( i.e., y_i \) and \( D \)).
Intrinsic geometric properties (i.e., $x_i$) of each neighborhood on $\mathcal{M}$ is equally valid for local patches on $\mathcal{N}$ [Roweis and Saul Science ’00].

Therefore, we can use the same set of local reconstruction codes $x_i$ to characterize the local geometric relationships between $g(y_i)$ and $g(D)$ on $\mathcal{N}$ as to characterize those between $y_i$ and $D$ on $\mathcal{M}$.
Learning Theory

Lemma

Let \( \mathcal{M}, \mathcal{N} \) and \( g \) be as above. Let \( p \in \mathcal{U}_p \) be an open subset of \( \mathcal{M} \) with respect to \( p \), such that \( \forall q \in \mathcal{U}_p \), the line segment \( \overline{pq} \) remains in \( \mathcal{U}_p \). If \( |\partial g^s/\partial q^t| \leq c, 1 \leq s \leq n, 1 \leq t \leq m \), at every \( q \in \mathcal{U}_p \), then we have \( \forall q \in \mathcal{U}_p \):

\[
\| g(q) - g(p) \|^2 \leq mnc^2 \| q - p \|^2.
\]

• Lemma indicates that as \( \mathcal{U}_p \) shrinks to be a sufficiently small neighborhood of \( p \), \( mnc^2 \| q - p \|^2 \rightarrow \| g(q) - g(p) \|^2 \rightarrow 0 \). We use this observation below.
**Learning Theory**

Our objective is to minimize \( \| g(y_i) - g(D)x_i \|_2 \) for all \( i \), which is equivalent to minimize \( \sum_{i=1}^{N} \| g(y_i) - g(D)x_i \|_2^2 \). Applying the previous Lemma, we derive the following theorem.

**Theorem**

Let \( g(y_i), y_i, g(D), D \) and \( g \) be as above. Let \( y_i \in U_{y_i} \) and \( D x_i \in U_{Dx_i} \) be open sets as in Lemma, that also satisfy \( D x_i \in U_{y_i} \) and \( \{ d_j | x_{ji} \neq 0, \forall j \} \subset U_{Dx_i}, \forall i \). If \( 1^Tx_i = 1 \) and \( \| x_i \|_0 = \tau \ (\tau \ll K) \) for all \( i \), then the following inequality holds:

\[
\sum_{i=1}^{N} \| g(y_i) - g(D)x_i \|_2^2 \leq \alpha \sum_{i=1}^{N} \| y_i - D x_i \|_2^2 + \beta \sum_{i=1}^{N} \sum_{j=1}^{K} \| x_{ji} \|_2^2 \leq \| D x_i - d_j \|_2^2
\]

where \( x_{ji} \) is the \( j \)-th element in vector \( x_i \), \( \tau \in \mathbb{Z}^+ \), and \( \alpha = 2c_1, \beta = 2\tau c_2 \), with \( c_1 = \sup(\{ |\partial g^s/\partial q^t| | q \in U_{y_i}, \forall i, s, t \} ) \), and \( c_2 = \sup(\{ |\partial g^s/\partial q^t| | q \in U_{Dx_i}, \forall i, s, t \} ) \). Note that \( i \) exclusively represents the indexes of \( y_i \) and its code \( x_i \) while \( j \) only denotes the \( j \)-th element in \( x_i \).
**Interpretation**

\[
\sum_{i=1}^{N} \| g(y_i) - g(D)x_i \|_2^2 \leq \alpha \sum_{i=1}^{N} \| y_i - Dx_i \|_2^2 + \beta \sum_{i=1}^{N} \sum_{j=1}^{K} \left[ x_{ji}^2 \| Dx_i - d_j \|_2^2 \right]
\]

- **approximation error**
- **localization error**

\[
\sum_{i=1}^{N} \| y_i - Dx_i \|_2^2 \approx \sum_{i=1}^{N} \sum_{j=1}^{K} \left[ x_{ji}^2 \| y_i - d_j \|_2^2 \right]
\]

- All \( d_j \in \{d_j|x_{ji} \neq 0, \forall j\} \rightarrow Dx_i \rightarrow y_i \), indicating that

\[
\beta \sum_{i=1}^{N} \sum_{j=1}^{K} \left[ x_{ji}^2 \| Dx_i - d_j \|_2^2 \right] \approx \beta \sum_{i=1}^{N} \sum_{j=1}^{K} \left[ x_{ji}^2 \| y_i - d_j \|_2^2 \right]
\]
Locality Constrained Dictionary Learning (LCDL)

Let $Y \in \mathbb{R}^{m \times N}$, $D \in \mathbb{R}^{m \times K}$, $X \in \mathbb{R}^{K \times N}$ be defined as above. We formulate the practical LCDL optimization problem as:

$$\min_{D, X} \|Y - DX\|_F^2 + \lambda \sum_{i=1}^{N} \sum_{j=1}^{K} \left[ x_{ji}^2 \|y_i - d_j\|_2^2 \right] + \mu \|X\|_F^2$$

s.t. \[
\begin{cases}
1^T x_i = 1 & \forall i \\
x_{ji} = 0 & \text{if } d_j \notin \Omega_\tau(y_i) \quad \forall i, j
\end{cases}
\]

where $\Omega_\tau(y_i)$ is defined as the $\tau$-neighborhood containing $\tau$ nearest neighbors of $y_i$. The sum-to-one constraint (*) follows from the symmetry requirement, while the locality constraint (**) allows $x_i$ to characterize the intrinsic local geometry.

- Minimizing the proposed LCDL problem yields a codebook of $K$ locality-preserving landmark points located on manifold $\mathcal{M}$ in observation space.
Outline

1. Introduction
   - Overview
   - NLDR and Examples
   - Motivations
   - Proposed Framework

2. Background Knowledge

3. Proposed Approach
   - Problem Formulation
   - Locality-Constrained Dictionary Learning

4. Experiments
   - Synthetic Datasets
   - Real-world Datasets

5. Conclusion and Future Work

6. Appendix
Experiments

The proposed LCDL algorithm is evaluated in two experimental scenarios

1. The effectiveness in approximating the intrinsic manifold
2. The performance in classification of the reconstructed low-dimensional manifold produced by LCDL

LCDL is compared with state-of-the-art dictionary learning algorithm K-SVD [Aahron et al. TSP ’06] and locality-preserving codebook learning algorithms Local Coordinate Coding (LCC) [Yu et al. NIPS ’09] and Locality-constrained Linear Coding (LLC) [Wang et al. CVPR ’10]
Experiment Setup

- Three synthetic manifolds are employed, *i.e.*, Swiss roll, Punctured sphere and Gaussian manifold.
- For each synthetic dataset, $N = 3000$ training data are randomly generated.
- We set $K = 500, 200, \text{ and } 100$ for the three manifold, respectively.
- The NLDR algorithms are Hessian LLE, Laplacian Eigenmap, and LLE for these three manifolds.
- Measure the root mean square error (RMSE) introduced through the reconstruction of an intrinsic manifold $\mathcal{N}$, *i.e.,*

$$\| g(Y) - g(D)X \|_F / \sqrt{N}$$

where $g(Y)$ and $g(D)$ are the low-dimensional embedding of training data and landmark points, respectively, computed via the NLDR algorithm.
Synthetic Datasets

Figure: Low-dimensional embedding reconstruction comparison on Swiss roll (1st row), Punctured sphere (2nd row) and Gaussian (3rd row). Ground truth means the low-dimensional embedding obtained directly from all training samples. The nearest neighbor parameter $k$ of NLDR algorithms is set to 6.
Face Recognition

- **Extended Yale B Database**
  - 38 persons, 2414 frontal face images of size $32 \times 32$
  - 32 images per person are randomly selected for training and the rest for testing

- **CMUPIE Database**
  - 68 persons, 11554 frontal face images of size $32 \times 32$
  - 130 images per person are randomly selected for training and the rest for testing
Experiment Setup

- The goal is to examine which dictionary learning algorithm yields the most meaningful low-dimensional embedding for classification.
- For all algorithms, a structured dictionary is learned as $D = [D_1 | D_2 | \ldots | D_C]$, where $D_i$ is the sub-dictionary for class $i$.
- The number of atoms per class is set to 8, yielding a dictionary of 304 atoms for the Extended YaleB Database and a dictionary of 544 atoms for the CMU PIE Database.
- All Train is selected as the baseline method, which represents the results obtained in performing LLE on the entire training set.
- Random is employed for comparison, meaning using randomly selected training samples as the dictionary.
- Nearest-neighbor classifier is employed.
Fix dictionary size 304 atoms for the Extended YaleB Database and 544 atoms for the CMU PIE Database

Vary the reduced dimension
Face Recognition

- Change dictionary size from 2 atoms per class to 10 atoms per class
- Fixed dimension

![Graphs showing recognition rate vs dictionary size for Extended YaleB and CMU PIE datasets.](image)

- **Extended YaleB**
  - Recognition Rate (%) vs Dictionary Size
  - Data points for LCDL, KSVD, LCC, and LLC

- **CMU PIE**
  - Recognition Rate (%) vs Dictionary Size
  - Data points for LCDL, KSVD, LCC, and LLC
Parameter and Performance

- Vary $\lambda$ among 0.001, 0.01, 0.1 and 1
- Vary $\tau$ among 1, 2, 3, 4 and 5

![Extended YaleB](image)

- ![CMU PIE](image)

Yin (Joe) Zhou (University of Delaware)  
April 5th 2013
Computational Cost

Table: The overall time (seconds) includes dictionary learning and training data embedding. Note the time measurement may vary based on different implementations.

<table>
<thead>
<tr>
<th></th>
<th>Extended YaleB</th>
<th>CMU PIE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall Time</td>
<td>Speedup</td>
</tr>
<tr>
<td>All Train</td>
<td>22.1577 s</td>
<td>1x</td>
</tr>
<tr>
<td>K-SVD</td>
<td>71.2387 s</td>
<td>0.3x</td>
</tr>
<tr>
<td>LCC</td>
<td>38.7172 s</td>
<td>0.6x</td>
</tr>
<tr>
<td>LLC</td>
<td>11.6593 s</td>
<td>1.9x</td>
</tr>
<tr>
<td>LCDL</td>
<td>7.1001 s</td>
<td>3.1x</td>
</tr>
</tbody>
</table>
Outline

1. Introduction
   - Overview
   - NLDR and Examples
   - Motivations
   - Proposed Framework

2. Background Knowledge

3. Proposed Approach
   - Problem Formulation
   - Locality-Constrained Dictionary Learning

4. Experiments
   - Synthetic Datasets
   - Real-world Datasets

5. Conclusion and Future Work

6. Appendix
Conclusion and Future Work

Conclusions:

- We show that the approximation to an unobservable intrinsic manifold by a few latent points residing on the manifold can be cast in a novel dictionary learning problem over the observation space.
- The presented locality constrained dictionary learning (LCDL) is a novel algorithm, which effectively learns a compact set of atoms consisting of locality-preserving landmark points on a nonlinear manifold.
- LCDL is superior to existing dictionary learning algorithms in terms of yielding more meaningful atoms for NLDR algorithms with greatly reduced computational complexity.

Future work includes:

- Testing over additional datasets
- Incorporating a sparse outlier term to improve robustness
- Extending LCDL to be a discriminative dictionary learning algorithm for classification
Outline

1 Introduction
   - Overview
   - NLDR and Examples
   - Motivations
   - Proposed Framework

2 Background Knowledge

3 Proposed Approach
   - Problem Formulation
   - Locality-Constrained Dictionary Learning

4 Experiments
   - Synthetic Datasets
   - Real-world Datasets

5 Conclusion and Future Work

6 Appendix
SHREC’ 11 Contest Dataset

- 30 classes, 600 watertight meshes
- 20 non-rigid shapes per class
Extension to 3D Object Recognition

Figure: Performance comparison on robustness against partial occlusion.
Proof Sketch

Denote by $Y \in \mathbb{R}^{m \times N}$ the matrix containing all $y_i$ and let $X = [x_1, \ldots, x_N] \in \mathbb{R}^{K \times N}$ be the matrix containing $N$ local reconstruction codes. We have

$$\sum_{i=1}^{N} \|g(y_i) - g(D)x_i\|_2^2 = \|g(Y) - g(D)X\|_F^2 \leq 2\|g(Y) - g(DX)\|_F^2 + 2\|g(DX) - g(D)X\|_F^2 \leq 2 \sum_{i=1}^{N} \|g(y_i) - g(Dx_i)\|_2^2 + 2 \sum_{i=1}^{N} \|g(Dx_i) - g(D)x_i\|_2^2$$

where in (a) $g(DX) \in \mathbb{R}^{n \times N}$ is a matrix representing the image of the reconstructed signals $DX$ via $g$; (b) is from Cauchy-Schwarz inequality; in (c) $g(Dx_i) \in \mathbb{R}^n$ is the $i$-th column in $g(DX)$. 
Proof Sketch

Since $1^T x_i = \sum_{j=1}^K x_{ji} = 1$ and $\|x_i\|_0 = \tau$ for all $i$, Eq. (1) can be written as:

$$\sum_{i=1}^N \|g(y_i) - g(D)x_i\|^2_2$$

$$\leq 2 \sum_{i=1}^N \|g(y_i) - g(Dx_i)\|^2_2 + 2 \sum_{i=1}^N \left\| \sum_{j=1}^K x_{ji} \left[ g(Dx_i) - g(d_j) \right] \right\|_2^2$$

$$\leq 2 \sum_{i=1}^N \|g(y_i) - g(Dx_i)\|^2_2 + 2\tau \sum_{i=1}^N \sum_{j=1}^K \left[ x_{ji}^2 \|g(Dx_i) - g(d_j)\|^2_2 \right]$$

Applying Lemma 1 to each $\|g(y_i) - g(Dx_i)\|^2_2$ and to each $x_{ji}^2 \|g(Dx_i) - g(d_j)\|^2_2$, $\exists c_1 = \sup (\{|\partial g^s / \partial q^t| \mid q \in U_{y_i}, \forall i, s, t\})$ and $c_2 = \sup (\{|\partial g^s / \partial q^t| \mid q \in U_{Dx_i}, \forall i, s, t\})$ such that $2\|g(y_i) - g(Dx_i)\|^2_2$

$$\leq 2c_1 \|y_i - Dx_i\|^2_2, \forall i \text{ and } 2\tau \left[ x_{ji}^2 \|g(Dx_i) - g(d_j)\|^2_2 \right] \leq 2\tau c_2 \left[ x_{ji}^2 \|Dx_i - d_j\|^2_2 \right], \forall i, j$$. Letting $\alpha = 2c_1$ and $\beta = 2\tau c_2$ completes the result.
Optimization

Block Coordinate Descent Method.

1. **Repeat**

2. for $i = 1$ to $N$ do
   Computing local reconstruction codes as
   \[
   \hat{x}_i \leftarrow \frac{(G + \lambda \delta(G) + \mu I)^{-1}1}{1^T(G + \lambda \delta(G) + \mu I)^{-1}1}
   \]
   end for

3. for $j = 1$ to $K$ do
   Updating dictionary atoms as
   \[
   d_j \leftarrow \frac{1}{(1 + \lambda)(x_{j*}x_{j*}^T)}(Ex_{j*}^T + \lambda Y \alpha)
   \]
   end for

4. **Until convergence**