# Dimensionality Reduction and Its Application in Data Mining

16 Nov 2017

CS5234 Mini Project

Yang Ruizhi, Chen Shaozhuang, Li Yuanda

# Agenda

1. Introduction to Dimensionality Reduction
2. Experiments with Dimensionality Reduction
   a. Text
   b. Image
3. Experiments with K-Means and K-NN
   a. Text
   b. Image
4. Conclusion

# Motivation for Dimensionality Reduction

- Many data mining/machine learning applications deal with high dimensional data
  - E.g. text, image, etc.
- Project high dimensional data to lower dimensional space is desirable
  - $R^d \rightarrow R^k$, where $k << d$

# Dimensionality Reduction Techniques

- We consider 3 DR techniques
  - Principle Component Analysis
  - (Gaussian) Random Projection
  - Sparse Random Projection

# Principle Component Analysis (PCA)

Let $X_{d \times N}$ be the original data set, which contains $N$ data points of dimension $d$. PCA requires **eigenvalue decomposition** of the data covariance matrix:

$$\frac{1}{N-1} X X^T = E \Lambda E^T$$

The $N$ $k$-dimensional data matrix after PCA is:

$$X_{PCA} = E_k^T X$$

Where $E_k$ contains $k$ eigenvectors corresponding to the $k$ largest eigenvalues in $\mathbf{\Lambda}$.

**Time Complexity**: $O(d^2N + d^3)$

# (Gaussian) Random Projection

Random projection is performed by simply multiply the original data matrix by a random matrix:

$$X_{RP} = RX$$

where the random matrix R is obtained by sampling each of entry from a Gaussian distribution *N(0, 1/k)*.

**Time Complexity**: *O(dkN)*

# Sparse Random Projection

Similar to Gaussian random projection, sparse random projection is done by simply multiply the original data matrix by a random matrix:

$$X_{RP}=RX$$

where the random matrix R is obtained by sampling each of entry using:

$$r_{ij}=\sqrt{3}\cdot\begin{cases}+1, & withprobability 1/6\\ 0, & withprobability 2/3\\ -1, & withprobability 1/6\end{cases}$$

**Time Complexity**: *O(dkN)*. But sparsity can be exploited.

# Johnson-Lindenstrauss Lemma (J-L Lemma)

There are various proofs and interpretations. Our way to interpret J-L Lemma:

> For a given error bound $\epsilon$ and a given size of data set $N$, as long as $k$ is suitably big, we can always find a random projection, such that the pairwise distance after projection is $\epsilon$-approximation of the original pairwise distance.

In human language:

- Pairwise distance between data points is nearly preserved.
- Performance guarantee for random projection.

# Summary

| | |
|---|---|
| **Random Projection (RP)** | $X_{RP}=RX$<br>X: original data matrix<br>R: projection matrix, Gaussian-sampled. |
| **Sparse Random Projection (SRP)** | $r_{ij}$ sampled from:<br>$r_{ij} = \sqrt{3}$ w.p. ⅙<br>$r_{ij} = 0$ w.p. ⅔<br>$r_{ij} = -\sqrt{3}$ w.p. ⅙ |
| **Principle Component Analysis (PCA)** | $X_{PCA}=E_K{}^T X$<br>X: original data matrix<br>$E_k$: contains k eigenvectores |

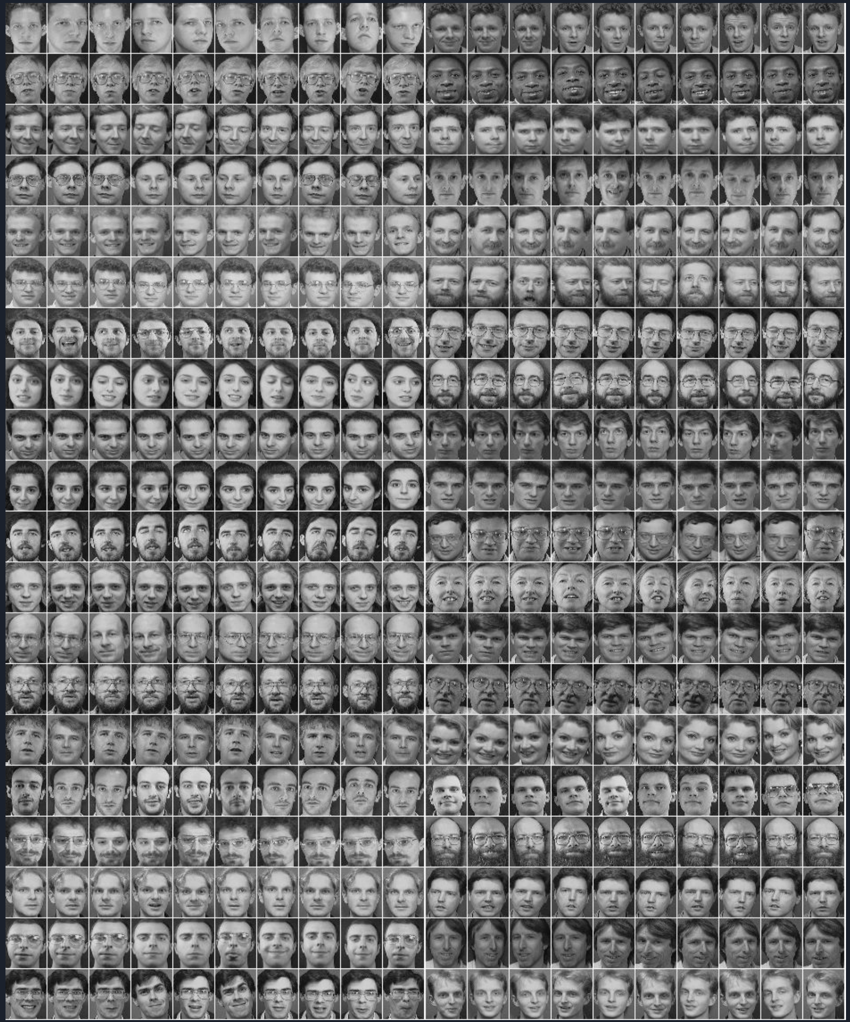
# Agenda

1. Introduction to Dimensionality Reduction
2. **Experiments with Dimensionality Reduction**
    a. **Text**
    b. **Image**
3. Experiments with K-Means and K-NN
    a. Text
    b. Image
4. Conclusion

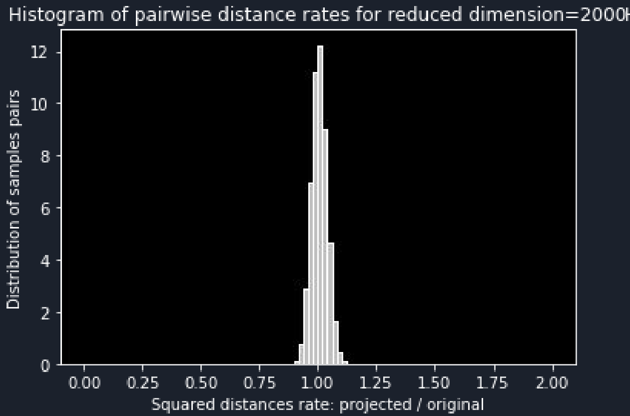
# Experimental Settings

- Objective: Justify the ability of different dimensionality reduction methods for preserving pairwise distance.
- Datasets (high-dimensional)
  - Text Data: 20 newsgroups dataset ($10^4$ dimension)
  - Image Data: Olivetti faces dataset ($10^3$ dimension)

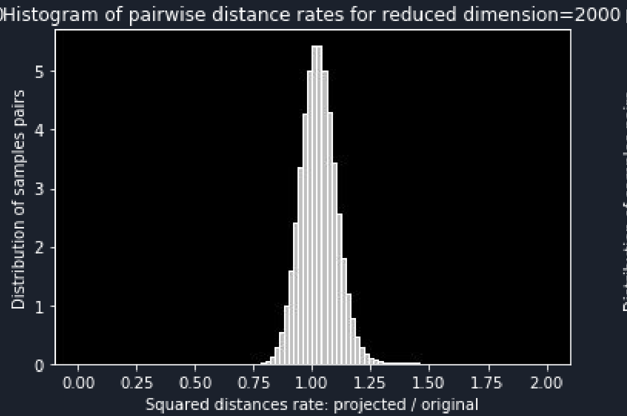
| | | |
|---|---|---|
| comp.graphics<br>comp.os.ms-windows.misc<br>comp.sys.ibm.pc.hardware<br>comp.sys.mac.hardware<br>comp.windows.x | rec.autos<br>rec.motorcycles<br>rec.sport.baseball<br>rec.sport.hockey | sci.crypt<br>sci.electronics<br>sci.med<br>sci.space |
| misc.forsale | talk.politics.misc<br>talk.politics.guns<br>talk.politics.mideast | talk.religion.misc<br>alt.atheism<br>soc.religion.christian |



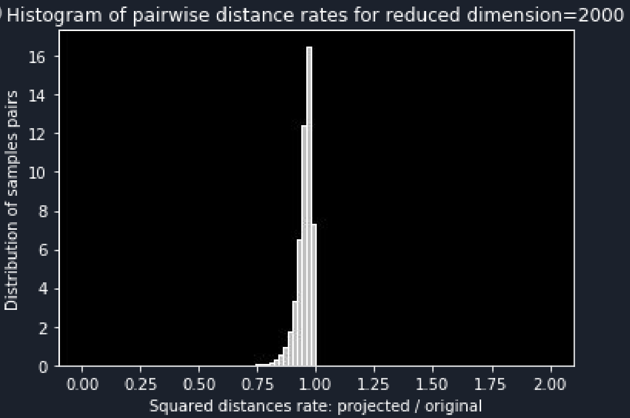
# Histogram for RP, SPR, PCA on Text Data




RP

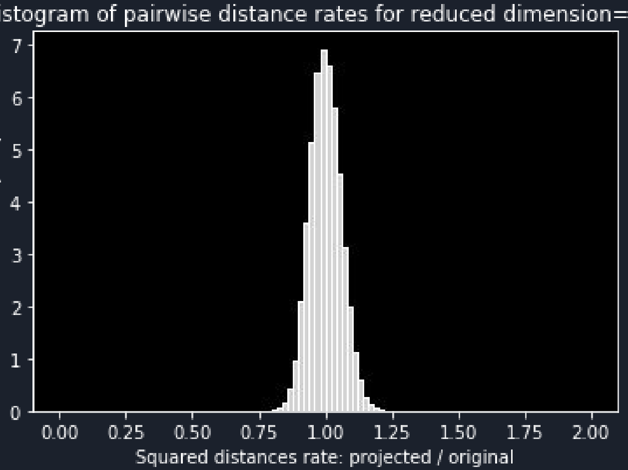



SRP

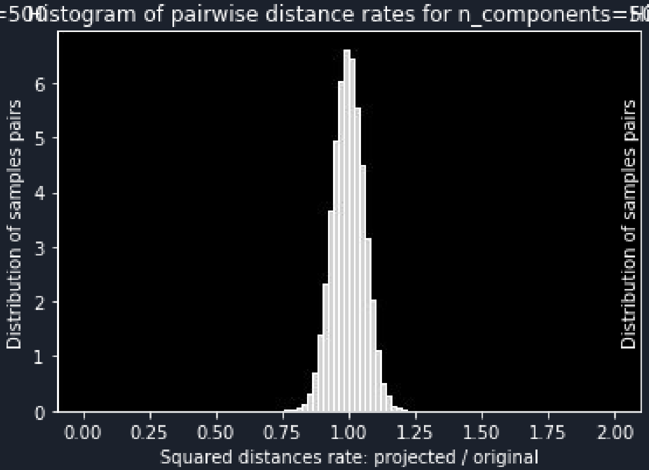



PCA

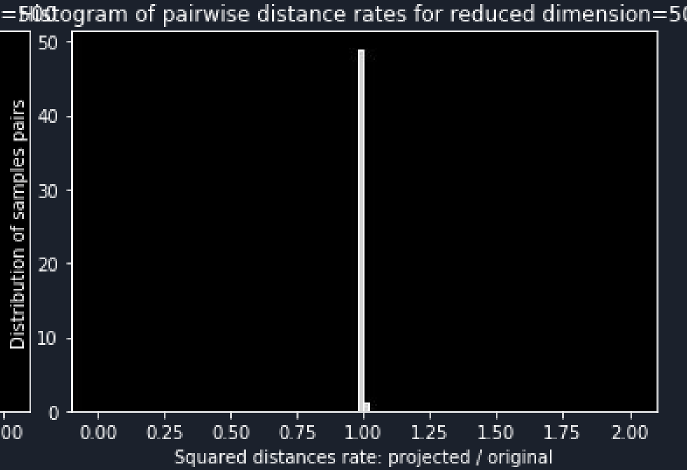
# Histogram for RP, SPR, PCA on Image Data




RP




SRP




PCA

# Agenda

1. Introduction to Dimensionality Reduction
2. Experiments with Dimensionality Reduction
    a. Text
    b. Image
3. **Experiments with K-Means and K-NN**
    a. **Text**
    b. **Image**
4. Conclusion

# Experimental Settings

We are particularly interested in data mining/machine learning algorithms that make heavy use of **pairwise distance** among data points:

- K-nearest neighbors (classification, supervised learning)
- K-means (clustering, unsupervised learning)

Number of categories:

- Text dataset: 5 categories considered.
- Image dataset: 10 categories considered.

# Performance Metrics

Training/test data division for kNN:

- Hold-out, 6 : 4. Stratified sampling.

Metrics:

- K-NN:
  - Precision, recall, f1-score
  - Higher means better.
- K-Means:
  - Homogeneity, completeness, v-measure.
  - Higher means better.

Each experiment is performed 20 times. Mean is taken for evaluation.

# K-Means on Text Data

| | Original Text + K-Means | RP + K-Means | SRP + K-Means | PCA + K-Means |
|---|---|---|---|---|
| Time | 77.56s | 5.6s +16.89s | 2.7s + 16.92s | 73s + 15.92s |
| Homogeneity | 0.467 | 0.470 | 0.465 | 0.461 |
| Completeness | 0.553 | 0.562 | 0.544 | 0.556 |
| V-measure | 0.506 | 0.517 | 0.501 | 0.504 |

# K-Means on Image Data

| | Original Image + K-Means | RP + K-Means | SRP + K-Means | PCA + K-Means |
|---|---|---|---|---|
| Time | 0.381s | 0.07s + 0.181s | 0.05s + 0.174s | 0.02s + 0.114s |
| Homogeneity | 0.646 | 0.639 | 0.599 | 0.580 |
| Completeness | 0.675 | 0.667 | 0.620 | 0.603 |
| V-measure | 0.660 | 0.653 | 0.609 | 0.591 |

# K-NN on Text Data

| | Original Image + K-NN | RP + K-NN | SRP + K-NN | PCA + K-NN |
|---|---|---|---|---|
| Time | 122.48s | 8.23s + 6.15s | 3.13s + 6.17s | 111.72s + 6.20s |
| Precision | 0.84 | 0.82 | 0.81 | 0.82 |
| Recall | 0.79 | 0.80 | 0.80 | 0.74 |
| F1-Score | 0.79 | 0.80 | 0.79 | 0.75 |

# K-NN on Image Data

| | Original Image + K-NN | RP + K-NN | SRP + K-NN | PCA + K-NN |
|---|---|---|---|---|
| Time | 0.381s | 0.07s + 0.181s | 0.05s + 0.174s | 0.02s + 0.114s |
| Precision | 0.815 | 0.797 | 0.799 | 0.814 |
| Recall | 0.715 | 0.719 | 0.721 | 0.718 |
| F1-Score | 0.714 | 0.714 | 0.714 | 0.713 |

# Agenda

1. Introduction to Dimensionality Reduction
2. Experiments with Dimensionality Reduction
    a. Text
    b. Image
3. Experiments with K-Means and K-NN
    a. Text
    b. Image
4. **Conclusion**

# Conclusion

Our results show:

- All three DR techniques perform well on our datasets and applications.
- Gaussian RP outperforms others in most of our experiments.
- PCA is inefficient on large dataset with large dimensionality.

We conclude:

- RP and SRP are good alternatives to PCA, when pairwise distance is important, dimension is big, and time complexity is a concern.

# References

[1] Johnson, W. B., & Lindenstrauss, J. (1984). Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, *26*(189-206), 1.

[2] Achlioptas, D. (2001, May). Database-friendly random projections. In *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 274-281). ACM.

[3] Dasgupta, S., & Gupta, A. (2003). An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, *22*(1), 60-65.