Dimensionality Reduction: Principle Component Analysis

Lingqiao Liu
What is dimensionality reduction?

\[ \mathbf{X} \in \mathbb{R}^{n \times p} \]
Dimensionality Reduction: why?

- Extract underlying factors

<table>
<thead>
<tr>
<th></th>
<th>Disagree</th>
<th>Neutral</th>
<th>Agree</th>
</tr>
</thead>
<tbody>
<tr>
<td>I am the life of the party.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel little concern for others.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am always prepared.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I get stressed out easily.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I have a rich vocabulary.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I don’t talk a lot.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am interested in people.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I leave my belongings around.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I am relaxed most of the time.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I have difficulty understanding abstract ideas.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I feel comfortable around people.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I insult people.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I pay attention to details.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I worry about things.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I have a vivid imagination.</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The five factors [edit]

A summary of the factors of the Big Five and their corresponding questions:

- **Openness to experience**: (inventive/curious vs. intellectual curiosity, creativity and a preference for a variety of activities over a strict routine)
- **Conscientiousness**: (efficient/organized vs. easygoing/spontaneous behavior)
- **Extraversion**: (outgoing/energetic vs. solitary/shy)
- **Agreeableness**: (friendly/compassionate vs. anxious/aloof)
- **Neuroticism**: (sensitive/nervous vs. secure/confident)
Dimensionality Reduction: why?

- Reduce data noise
  - Face recognition
  - Applied to image de-noising

Image courtesy of Charles-Alban Deledalle, Joseph Salmon, Arnak Dalalyan; BMVC 2011

Image denoising with patch-based PCA: local versus global
Dimensionality Reduction: why?

- Reduce the number of model parameters
  - Avoid over-fitting
  - Reduce the computational load
Dimensionality Reduction: why?

- Visualization
Dimensionality Reduction

- **General principle:**
  - Preserve “useful” information in low dimensional data

- **How to define “usefulness”?**
  - Many
  - An active research direction in machine learning

- **Taxonomy**
  - Supervised or Unsupervised
  - Linear or nonlinear

- **Commonly used methods:**
  - PCA, LDA (linear discriminant analysis), local linear embedding and more.
Outline

- Theoretic Part
  - PCA explained: two perspectives
  - Mathematic basics
  - PCA objective and solution

- Application Example
  - Eigen face
  - Handle high feature dimensionality

- Extensions:
  - Kernel PCA
PCA explained: Two perspectives

- Data correlation and information redundancy
- Signal-noise ratio maximization
**PCA explained: Data correlation and information redundancy**

<table>
<thead>
<tr>
<th>Statement</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>I have a rich vocabulary.</td>
<td></td>
</tr>
<tr>
<td>I don't talk a lot.</td>
<td></td>
</tr>
<tr>
<td>I am interested in people.</td>
<td></td>
</tr>
<tr>
<td>I leave my belongings around.</td>
<td></td>
</tr>
<tr>
<td><strong>I am relaxed most of the time.</strong></td>
<td></td>
</tr>
<tr>
<td>I have difficulty understanding abstract ideas.</td>
<td></td>
</tr>
<tr>
<td>I feel comfortable around people.</td>
<td></td>
</tr>
<tr>
<td>I insult people.</td>
<td></td>
</tr>
<tr>
<td>I pay attention to details.</td>
<td></td>
</tr>
<tr>
<td><strong>I worry about things.</strong></td>
<td></td>
</tr>
<tr>
<td>I have a vivid imagination.</td>
<td></td>
</tr>
</tbody>
</table>
PCA explained: Data correlation and information redundancy

- Dependency vs. Correlation
  - Dependent is a stronger criterion

- Equivalent when data follows Gaussian distribution

- PCA only de-correlates data
  - One limitation of PCA
  - ICA, but it is more complicated
PCA explained: Data correlation and information redundancy

- Geometric interpretation of correlation
PCA explained: Data correlation and information redundancy
PCA explained: Data correlation and information redundancy
PCA explained: Data correlation and information redundancy

- Correlation can be removed by rotating the data point or coordinate.
PCA explained: Signal-noise ratio maximization

- Maximize

\[ SNR = \frac{\sigma^2_{signal}}{\sigma^2_{noise}}. \]
PCA explained: Signal-noise ratio maximization

- Keep one signal dimension, discard one noisy dimension
PCA explained: Signal-noise ratio maximization
PCA explained

- **Target**
  - 1: Find a new coordinate system which makes different dimensions zero correlated
  - 2: Find a new coordinate system which aligns (top-k) largest variance

- **Method**
  - Rotate the data point or coordinate

- **Mathematically speaking…**
  - How to rotate?
  - How to express our criterion
Mathematic Basics

- Mean, Variance, Covariance
- Matrix norm, trace,
- Orthogonal matrix, basis
- Eigen decomposition
Mathematic Basics

- (Sample) Mean

\[ \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \]

- (Sample) Variance

\[ s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1} \]

- (Sample) Covariance

\[ cov(X, Y) = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1} \]
Mathematic Basics

- Covariance Matrix

\[
C = \begin{pmatrix}
cov(x, x) & cov(x, y) & cov(x, z) \\
cov(y, x) & cov(y, y) & cov(y, z) \\
cov(z, x) & cov(z, y) & cov(z, z)
\end{pmatrix}
\]

\[
C = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^T
\]
Mathematic Basics

- Frobenius norm

\[ \|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} \]

- Trace

\[ \text{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn} = \sum_{i=1}^{n} a_{ii} \]

\[ \text{tr}(X^T Y) = \text{tr}(X Y^T) = \text{tr}(Y^T X) = \text{tr}(Y X^T) \]

\[ \|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\text{trace}(A^* A)} \]
Mathematic Basics

- Symmetric Matrix \( A = A^T \)
- Covariance matrix is symmetric

\[
C = \frac{1}{n - 1} XX^T
\]

\[
C = C^T
\]
Mathematic Basics

- Orthogonal matrix

\[ Q^T Q = QQ^T = I \]

- Rotation effect

\[ \|Qx\|_F = \sqrt{\text{trace}(x^T Q^T Qx)} = \sqrt{\text{trace}(x^T x)} = \|x\|_F \]

\[ x = Q^T Q x \]
Mathematic Basics

- Relationship to coordinate system
  - A point = linear combination of bases
  - Combination weight = coordinate
- Each row (column) of $Q = basis$
  - Not unique
  - Relation to coordinate rotation
- New coordinate $Qx$
Mathematic Basics

\[
\begin{pmatrix} 2 \\ 2 \end{pmatrix} = 2 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + 2 \begin{pmatrix} 0 \\ 1 \end{pmatrix}
\]

\[
\begin{pmatrix} 2 \\ 2 \end{pmatrix} = 2\sqrt{2} \begin{pmatrix} 1 \\ \sqrt{2} \end{pmatrix} + 0 \begin{pmatrix} 1 \\ -\sqrt{2} \end{pmatrix}
\]

New coordinate

\[
\begin{pmatrix} 2\sqrt{2} \\ 0 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ \sqrt{2} & -\sqrt{2} \end{pmatrix} \begin{pmatrix} 2 \\ 2 \end{pmatrix}
\]

Old coordinate
Mathematic Basics

• Eigenvalue and Eigenvector

\[ Au = \lambda u \]

• Properties
  ◦ Multiple solutions
  ◦ Scaling invariant
  ◦ Relation to the rank of A
Mathematic Basics

Eigen-decomposition

- If $A$ is symmetric

$$A = Q \Lambda Q^T \quad Q^T Q = QQ^T = I$$
PCA: solution

- Target 1: de-correlation

\[
C_X = \frac{1}{n-1} XX^T
\]

\[
Y = PX
\]

\[
C_Y = \frac{1}{n-1} PX(PX)^T
\]

\[
= \frac{1}{n-1} PXX^TP^T
\]

\[
= \frac{1}{n-1} PQ\Lambda Q^T P^T
\]
PCA: solution

\[ C_Y = \frac{1}{n-1} P X (P X)^T \]

\[ = \frac{1}{n-1} P X X^T P^T \]

\[ = \frac{1}{n-1} P Q \Lambda Q^T P^T \]

if \( P = Q^T \)

\[ C_Y = \frac{1}{n-1} \Lambda \]
PCA: solution

- Variance of each dimension

\[ \text{Var}(y_k) = \frac{1}{n-1} P_k X X^T P_k^T \]

\[ = \frac{1}{n-1} p_k Q \Lambda Q^T p_k^T \]

\[ = \frac{1}{n-1} \lambda_k \]

- Rank dimensions according to their corresponding eigenvalues
PCA: algorithm

1. Subtract mean
2. Calculate the covariance matrix
3. Calculate eigenvectors and eigenvalues of the covariance matrix
4. Rank eigenvectors by its corresponding eigenvalues
4. Obtain P with its column vectors corresponding to the top k eigenvectors
PCAS: MATLAB code

5  \begin{align}
6 & \text{Mu = mean(fea);} \\
7 & \text{fea = fea - repmat(Mu, [size(fea, 1), 1]);} \\
8 & \text{Cov = fea' * fea;} \\
9 & \text{[V, D] = eig(Cov);} \\
10 & \text{[value, rank_idx] = sort(diag(D), 'descend');} \\
11 & \text{P = V(:, rank_idx(1:10));}
\end{align}
PCA: reconstruction

- Reconstruct \( x \)

\[
\hat{x} = \hat{P}^T \hat{P} x
\]

- Derive PCA through minimizing the reconstruction error

\[
\min_{\hat{P}} \| X - \hat{P}^T \hat{P} X \|_F^2 \quad s.t. \quad \hat{P} \hat{P}^T = I
\]
PCA: reconstruction

- Reighley Quotient

\[
\max_P \text{trace}(PXX^T P^T)
\]

s.t. \( PP^T = I \)

- Solution = PCA
Application: Eigen-face method

- Sirovich and Kirby (1987) showed that PCA could be used on a collection of face images to form a set of basis features.
- Not only limited to face recognition
- Steps
  - Image as high-dimensional feature
  - PCA
Application: Eigen-face method

Some eigenfaces from AT&T Laboratories Cambridge
Application: Reconstruction

Reconstructed from top-2 eigenvectors
Application: Reconstruction

- Reconstructed from top-15 eigenvectors
Application: Reconstruction

- Reconstructed from top-40 eigenvectors
Application: Eigen-face method

- From large to small eigenvalues

- Common Patterns
- Discriminative Patterns
- Noisy Patterns
Eigen-face method: How to handle high dimensionality

- For high-dimensional data, \( C_X = \frac{1}{n-1} XX^T \) can be too large.
- The number of samples is relatively small:

\[
d \gg N
\]

\[
X^T X v = \lambda v \\
XX^T (X v) = \lambda (X v)
\]

Define \( u = X v \) \[
XX^T u = \lambda u
\]
Eigen-face method: How to handle high dimensionality

1. Centralize data
2. Calculate the kernel matrix
3. Perform Eigen-decomposition on the kernel matrix and obtain its eigenvector $\mathbf{V}$
4. Obtain the Eigenvector of the covariance matrix by $\mathbf{u} = \mathbf{Xv}$

Question? How many eigenvectors you can obtain in this way?
Extension: Kernel PCA

- Kernel Method

\[ x \rightarrow \phi(x) \]

- You do not have access to \( \phi(x) \) but you know

\[ k(x, y) = \langle \phi(x), \phi(y) \rangle \]
Extension: Kernel PCA

- Use the same way as how we handle high-dimensional data

\[ u = \phi(X)v \]

\[ u^T x^* = (\phi(X)v)^T x_i = \sum_{j=1}^{N} v_j \langle \phi(x_j), \phi(x^*) \rangle = \sum_{j=1}^{N} v_j k(x_j, x^*) \]