Dimensionality Reduction with PCA

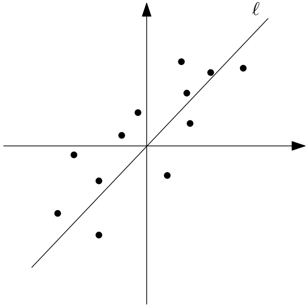
Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

Dimensionality Reduction

Let P be a set of n points in d-dimensional space, where d is a very large value (possibly even larger than n). Informally, the goal of dimensionality reduction is to convert P into a set P' of points in a k-dimensional space where $k < d$, such that P' loses as little information about P as possible.

Example. We can convert 2d points into 1d ones by projecting them onto a line ℓ.



Why Dimensionality Reduction?

- Better mining efficiency and/or effectiveness.
  - Most data mining algorithms work poorly in high dimensional space (a phenomenon known as the curse of dimensionality).
- Compression.
- Data visualization.
- ...

Basics 1

- A vector $\boldsymbol{v}$ is a $d \times 1$ matrix: $\boldsymbol{v} = (v[1], ..., v[d])^T$.
- A point can be represented as vector.
- A vector $\boldsymbol{v}$ is a unit vector if $\sum_{i=1}^{d} v[i]^2 = 1$.
- Dot product $\boldsymbol{v_1} \cdot \boldsymbol{v_2} = \sum_{i=1}^{d}(v_1[i]v_2[i])$.
- If two vectors $\boldsymbol{v_1}, \boldsymbol{v_2}$ are orthogonal, $\boldsymbol{v_1} \cdot \boldsymbol{v_2} = 0$.
- Let $\boldsymbol{p}$ be a point and $\boldsymbol{v}$ a unit vector. Then, $\boldsymbol{p} \cdot \boldsymbol{v}$ gives the distance from the origin to the projection of $\boldsymbol{p}$ on $\boldsymbol{v}$.

Basics 2

Let S be a set of real numbers $r_1, ..., r_m$. The mean of S equals:

$$mean(S) = \frac{1}{m}\sum_{i=1}^{m} r_i.$$

The variance of S equals:

$$var(S) = \frac{1}{m}\sum_{i=1}^{m}(r_i - mean(S))^2.$$

Basics 3

Let P be a set of 2d points $p_1, ..., p_n$. Its co-variance between dimensions i and j (where $1 \leq i \leq j \leq d$) equals

$$cov = \frac{1}{n} \sum_{k=1}^{n} (p_k[i] - mean_i)(p_k[j] - mean_j)$$

where $mean_i$ ($mean_j$) is the mean of the coordinates in P along dimension i (j).

Basics 4

The co-variance matrix A of point set P is a $d \times d$ matrix whose value at the i-th row and j-th column ($i, j \in [1, d]$) is the co-variance of P between dimensions i and j.

Note that A is symmetric, namely, $A = A^T$.

Basics 5

Let A be a $d \times d$ matrix. If for some real value $d \times 1$ unit vector $\boldsymbol{v}$, it holds that

$$A\boldsymbol{v} = \lambda\boldsymbol{v}$$

then v is called a unit eigenvector of A, and λ is called an eigenvalue of A.

Principle Component Analysis (PCA)

algorithm (P, k)
/* output: $k \leq d$ directional vectors */

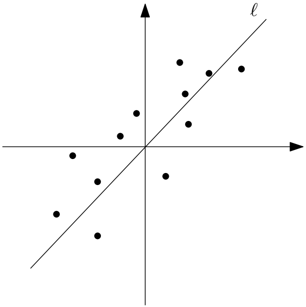
1. shift P such that its geometric mean is at the origin of the data space
2. $A \leftarrow$ the co-variance matrix of P
3. compute all the d unit eigenvectors
4. arrange the eigenvectors in descending order of their eigenvalues
5. return the first k eigenvectors $\boldsymbol{v_1}, ..., \boldsymbol{v_k}$

Note

Each point $\boldsymbol{p}$ is then converted to a k-dimensional point whose i-th $(1 \leq i \leq d)$ coordinate is $\boldsymbol{v_i} \cdot \boldsymbol{p}$.

Property of PCA

$\boldsymbol{v_1}$ is the direction along which the projections of P have the largest variance. In general, $\boldsymbol{v_i}$ $(i > 1)$ is the direction along which P has the largest variance, among all directions orthogonal to all of $\boldsymbol{v_1}, \ldots, \boldsymbol{v_{i-1}}$.



Next we will prove this fact for $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$. Then, the case with $\boldsymbol{v_3}, \ldots, \boldsymbol{v_i}$ follows the same idea.

Formally, let P be a set of n d-dimensional points with zero mean on all dimensions. Let $\boldsymbol{w}$ be a unit vector. We can project P onto $\boldsymbol{w}$ to obtain a set of 1d values: $S = \{\boldsymbol{p} \cdot \boldsymbol{w} \mid p \in P\}$. Define the quality of $\boldsymbol{w}$ be $var(S)$.

Theorem 1

The first eigenvector output by PCA has the highest quality.

Proof of Theorem 1

Let $\boldsymbol{X}$ be the $n \times d$ matrix where each row lists out the coordinates of a point in P. Thus, we can view S as a vector $\boldsymbol{Xw}$. Thus:

$$
\begin{aligned}
var(S) &= \frac{1}{n}(\boldsymbol{Xw})^T(\boldsymbol{Xw}) \\
&= \boldsymbol{w}^T \frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\boldsymbol{w} \\
&= \boldsymbol{w}^T\boldsymbol{Aw}
\end{aligned}
$$

where $\boldsymbol{A}$ is the covariance matrix of P. Hence, we want to maximize the above subject to the constraint that $\boldsymbol{w}^T\boldsymbol{w} = 1$.

Proof of Theorem 1 (Cont.)

Now we apply the method of Lagrange multipliers to find the maximum. Introduce a real value λ, and now consider the objective function

$$\begin{aligned} f(\boldsymbol{w}, \lambda) &= \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{w} - 1) \Rightarrow \\ \frac{\partial f}{\partial \boldsymbol{w}} &= 2\boldsymbol{A}\boldsymbol{w} - 2\lambda \boldsymbol{w} \end{aligned}$$

Equating the above 0 gives $\boldsymbol{A}\boldsymbol{w} = \lambda \boldsymbol{w}$. In other words, $\boldsymbol{w}$ needs to be an eigenvector, and λ the corresponding eigenvalue.

Proof of Theorem 1 (Cont.)

Now it remains to check which eigenvector gives the largest variance. Observe that:

$$\begin{aligned} var(S) &= \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w} \\ &= \boldsymbol{w}^T \lambda \boldsymbol{w} \\ &= \lambda \end{aligned}$$

In other words, when we choose eigenvector $\boldsymbol{w}$ as our solution, its quality is exactly the eigenvalue λ. Hence, the eigenvector with the maximum eigenvalue is what we are looking for. □

Recall our earlier definitions. P is a set of n d-dimensional points with zero mean on all dimensions. Let $\boldsymbol{w}$ be a unit vector. Project P onto $\boldsymbol{w}$ to obtain a set of 1d values: $S = \{\boldsymbol{p} \cdot \boldsymbol{w} \mid p \in P\}$. Define the quality of $\boldsymbol{w}$ be $var(S)$.

Theorem 2

The second eigenvector output by PCA has the highest quality, among all the vectors $\boldsymbol{w}$ orthogonal to the first eigenvector $\boldsymbol{v_1}$.

Proof of Theorem 2

Let $\boldsymbol{A}$ be the covariance matrix of P. As shown in the proof of Theorem 1, we proved that

$$var(S) \quad = \quad \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w}.$$

Hence, we want to maximize the above subject to the constraints $\boldsymbol{w}^T \boldsymbol{w} = 1$ and $\boldsymbol{w}^T \boldsymbol{v_1} = 0$.

Now we apply the method of Lagrange multipliers to find the maximum. Introduce real values λ and ϕ, and now consider the objective function

$$\begin{aligned} f(\boldsymbol{w}, \lambda, \phi) &= \boldsymbol{w}^T \boldsymbol{A} \boldsymbol{w} - \lambda(\boldsymbol{w}^T \boldsymbol{w} - 1) - \phi \boldsymbol{w}^T \boldsymbol{v_1} \Rightarrow \\ \frac{\partial f}{\partial \boldsymbol{w}} &= 2\boldsymbol{A}\boldsymbol{w} - 2\lambda \boldsymbol{w} - \phi \boldsymbol{v_1}. \end{aligned}$$

Proof of Theorem 2 (Cont.)

The optimal $\boldsymbol{w}$ needs to satisfy $\frac{\partial f}{\partial \boldsymbol{w}} = 0$, namely:

$$2\boldsymbol{A}\boldsymbol{w} - 2\lambda\boldsymbol{w} - \phi\boldsymbol{v_1} = 0. \tag{1}$$

Next we prove that ϕ must be 0. To see this, multiplying both sides of (1) by $\boldsymbol{v_1}^T$, we get:

$$2\boldsymbol{v_1}^T\boldsymbol{A}\boldsymbol{w} - 2\lambda\boldsymbol{v_1}^T\boldsymbol{w} + \phi\boldsymbol{v_1}^T\boldsymbol{v_1} = 0. \tag{2}$$

We know that $\boldsymbol{v_1}^T\boldsymbol{w} = 0$, and $\boldsymbol{v_1}^T\boldsymbol{v_1} = 1$. Furthermore,

$$\boldsymbol{v_1}^T\boldsymbol{A}\boldsymbol{w} = \boldsymbol{w}^T\boldsymbol{A}^T\boldsymbol{v_1} = \boldsymbol{w}^T\boldsymbol{A}\boldsymbol{v_1} = \boldsymbol{w}^T(\boldsymbol{A}\boldsymbol{v_1}) = \boldsymbol{w}^T\boldsymbol{v_1} = 0.$$

Hence, from (2), we get $\phi = 0$.

Proof of Theorem 2 (Cont.)

Therefore, from (1), we know:

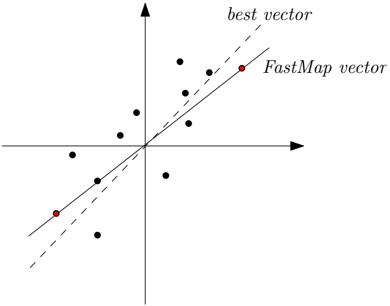
$$2\boldsymbol{A}\boldsymbol{w} - 2\lambda\boldsymbol{w} \quad = \quad 0$$

namely, $\boldsymbol{w}$ must also be an eigenvector.

From the proof of Theorem 1, we know that $var(S)$ equals the eigenvalue corresponding to $\boldsymbol{w}$. This thus indicates that $\boldsymbol{w}$ is the eigenvector of A with the second largest eigenvalue. □

When d is large, PCA is slow because it has to deal with a gigantic matrix with d^2 values. This motivates FastMap, which can be regarded as a heuristic version of PCA that trades precision for efficiency.

Heuristic 1 of FastMap. Assume that the vector between the farthest pair of points in P captures a large amount of variance of P.



Heuristic 2 of FastMap. Use the following algorithm to find the farthest pair of points in P

1. $p_1 \leftarrow$ a random point in P
2. $p_2 \leftarrow$ farthest point from p_1
3. $p_1 \leftarrow$ farthest point from p_2

FastMap

algorithm (P, k)
/* output: $k \leq d$ directional vectors */

1. for $i = 1$ to k
2. find the pair (p_1, p_2) of farthest points in P (Heuristic 2)
3. $\boldsymbol{v_i} = \boldsymbol{p_2} - \boldsymbol{p_1}$ (vector subtraction)
4. $P' \leftarrow$ the point set obtained by projecting the points in P onto the plane perpendicular to v_i
5. $P \leftarrow P'$
6. return $v_1, ..., v_k$