

Dimensionality Reduction with PCA

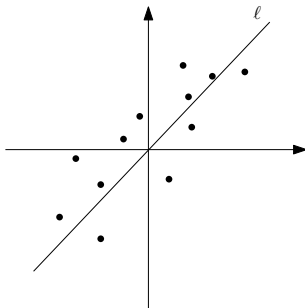
Yufei Tao

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.
- ...

- A **vector** \mathbf{v} is a $d \times 1$ matrix: $\mathbf{v} = (v[1], \dots, v[d])^T$.
- A point can be represented as vector.
- A vector \mathbf{v} is a **unit vector** if $\sum_{i=1}^d v[i]^2 = 1$.
- Dot product $\mathbf{v}_1 \cdot \mathbf{v}_2 = \sum_{i=1}^d (v_1[i]v_2[i])$.
- If two vectors $\mathbf{v}_1, \mathbf{v}_2$ are orthogonal, $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$.
- Let \mathbf{p} be a point and \mathbf{v} a unit vector. Then, $\mathbf{p} \cdot \mathbf{v}$ gives the distance from the origin to the projection of \mathbf{p} on \mathbf{v} .

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

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

The **variance** of S equals:

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

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

$$\text{cov} = \frac{1}{n} \sum_{k=1}^n (p_k[i] - \text{mean}_i)(p_k[j] - \text{mean}_j)$$

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

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, $A = A^T$.

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

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

then \mathbf{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 */

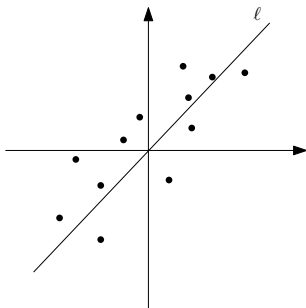
1. shift P such that its geometric mean is at the origin of the data space
2. $A \leftarrow$ the co-variance matrix of P
3. compute all the d unit eigenvectors
4. arrange the eigenvectors in **descending** order of their eigenvalues
5. return the first k eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_k$

Note

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

Property of PCA

\mathbf{v}_1 is the direction along which the projections of P have the largest variance. In general, \mathbf{v}_i ($i > 1$) is the direction along which P has the largest variance, among all directions orthogonal to all of $\mathbf{v}_1, \dots, \mathbf{v}_{i-1}$.



Next we will prove this fact for \mathbf{v}_1 and \mathbf{v}_2 . Then, the case with $\mathbf{v}_3, \dots, \mathbf{v}_i$ follows the same idea.

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

Theorem 1

The first eigenvector output by PCA has the highest quality.

Proof of Theorem ??

Let \mathbf{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 $\mathbf{X}\mathbf{w}$. Thus:

$$\begin{aligned} \text{var}(S) &= \frac{1}{n}(\mathbf{X}\mathbf{w})^T(\mathbf{X}\mathbf{w}) \\ &= \mathbf{w}^T \frac{\mathbf{X}^T \mathbf{X}}{n} \mathbf{w} \\ &= \mathbf{w}^T \mathbf{A} \mathbf{w} \end{aligned}$$

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

Proof of Theorem ?? (Cont.)

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

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

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

Proof of Theorem ?? (Cont.)

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

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

In other words, when we choose eigenvector \mathbf{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 \mathbf{w} be a unit vector. Project P onto \mathbf{w} to obtain a set of 1d values: $S = \{\mathbf{p} \cdot \mathbf{w} \mid \mathbf{p} \in P\}$. Define the **quality** of \mathbf{w} be $\text{var}(S)$.

Theorem 2

The second eigenvector output by PCA has the highest quality, among all the vectors \mathbf{w} orthogonal to the first eigenvector \mathbf{v}_1 .

Proof of Theorem ??

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

$$\text{var}(S) = \mathbf{w}^T \mathbf{A} \mathbf{w}.$$

Hence, we want to maximize the above subject to the constraints $\mathbf{w}^T \mathbf{w} = 1$ and $\mathbf{w}^T \mathbf{v}_1 = 0$.

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

$$\begin{aligned} f(\mathbf{w}, \lambda, \phi) &= \mathbf{w}^T \mathbf{A} \mathbf{w} - \lambda(\mathbf{w}^T \mathbf{w} - 1) - \phi \mathbf{w}^T \mathbf{v}_1 \Rightarrow \\ \frac{\partial f}{\partial \mathbf{w}} &= 2\mathbf{A} \mathbf{w} - 2\lambda \mathbf{w} - \phi \mathbf{v}_1. \end{aligned}$$

Proof of Theorem ?? (Cont.)

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

$$2\mathbf{A}\mathbf{w} - 2\lambda\mathbf{w} - \phi\mathbf{v}_1 = 0. \quad (1)$$

Next we prove that ϕ must be 0. To see this, multiplying both sides of (??) by \mathbf{v}_1^T , we get:

$$2\mathbf{v}_1^T \mathbf{A}\mathbf{w} - 2\lambda\mathbf{v}_1^T \mathbf{w} + \phi\mathbf{v}_1^T \mathbf{v}_1 = 0. \quad (2)$$

We know that $\mathbf{v}_1^T \mathbf{w} = 0$, and $\mathbf{v}_1^T \mathbf{v}_1 = 1$. Furthermore,

$$\mathbf{v}_1^T \mathbf{A}\mathbf{w} = \mathbf{w}^T \mathbf{A}^T \mathbf{v}_1 = \mathbf{w}^T \mathbf{A}\mathbf{v}_1 = \mathbf{w}^T (\mathbf{A}\mathbf{v}_1) = \mathbf{w}^T \mathbf{v}_1 = 0.$$

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

Proof of Theorem ?? (Cont.)

Therefore, from (??), we know:

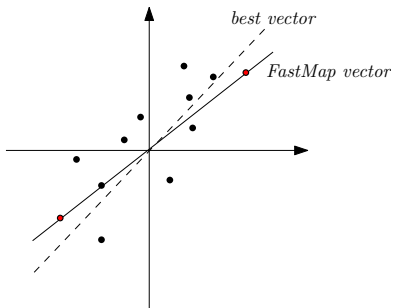
$$2\mathbf{A}\mathbf{w} - 2\lambda\mathbf{w} = 0$$

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

From the proof of Theorem ??, we know that $\text{var}(S)$ equals the eigenvalue corresponding to \mathbf{w} . This thus indicates that \mathbf{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

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. $\mathbf{v}_i = \mathbf{p}_2 - \mathbf{p}_1$ (vector subtraction)
4. $P' \leftarrow$ the point set obtained by projecting the points in P
 onto the plane perpendicular to \mathbf{v}_i
5. $P \leftarrow P'$
6. return $\mathbf{v}_1, \dots, \mathbf{v}_k$