Dimensionality Reduction with PCA

Yufei Tao

Department of Computer Science and Engineering
Chinese University of Hong Kong

P = a set of points in $\mathbb{R}^d$, where the dimensionality d is large.

The goal of **dimensionality reduction** is to convert P into a set P' of points in a lower-dimensional subspace such that P' does not lose "too much" information about P.

We will learn a classical method called **principled component analysis** (PCA) to achieve the purpose.

Subspace

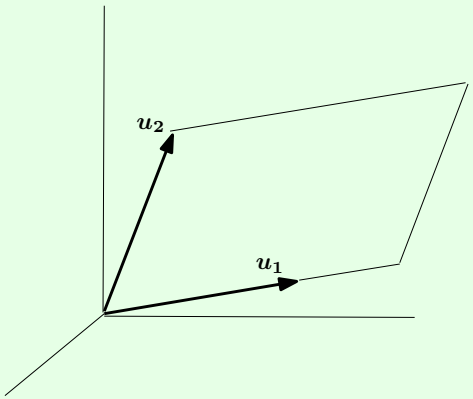
Fix an integer $k \leq d$.

A **k-subspace** Σ is the span of k unit vectors $\boldsymbol{u}_1, \boldsymbol{u}_2, ..., \boldsymbol{u}_k$ in $\mathbb{R}^d$ that are mutually orthogonal. We refer to $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_k\}$ as a **basis** of Σ.

Recall that the span of $\boldsymbol{u}_1, ..., \boldsymbol{u}_k$ is defined as

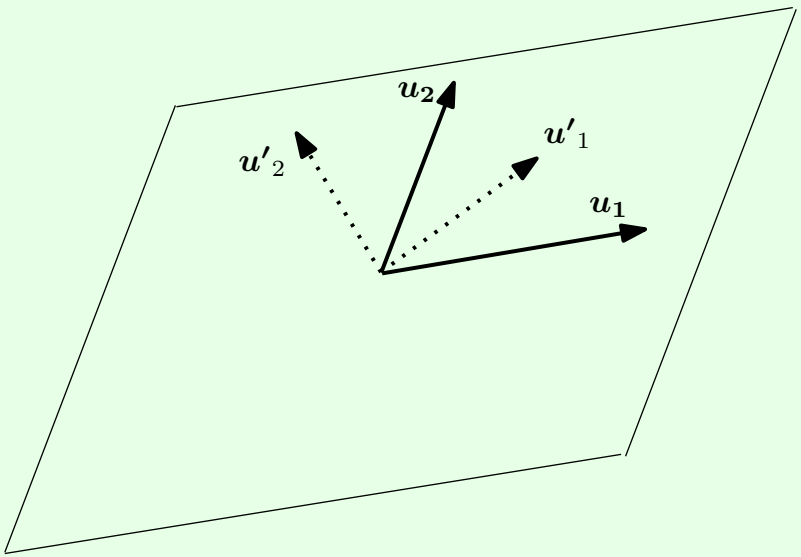
$$\left\{ \Sigma_{i=1}^{k} c_i \boldsymbol{u}_i \mid \text{ any real values } c_1, c_2, ..., c_k \right\}.$$

The span of two vectors is a plane:



Subspace

A k-subspace with $k \geq 2$ has an infinite number of bases.

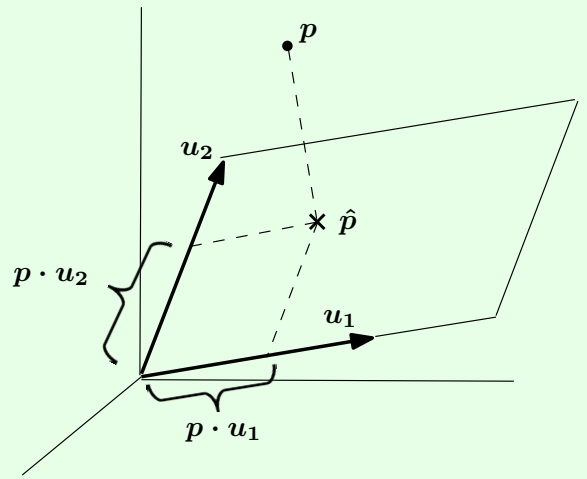


Projection

Let $\boldsymbol{p}$ be a point (a.k.a., a vector) in $\mathbb{R}^d$.
Consider a k-subspace Σ with a basis $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_k\}$.
For each $i \in [1, k]$, the **coordinate** of $\boldsymbol{p}$ on $\boldsymbol{u}_i$ is $\boldsymbol{p} \cdot \boldsymbol{u}_i$.
The **projection** of $\boldsymbol{p}$ onto Σ is

$$\hat{\boldsymbol{p}} = \Sigma_{i=1}^{k} c_i \boldsymbol{u}_i.$$

where $c_i = \boldsymbol{p} \cdot \boldsymbol{u}_i$. Note: $\hat{\boldsymbol{p}}$ is a d-dimensional point.

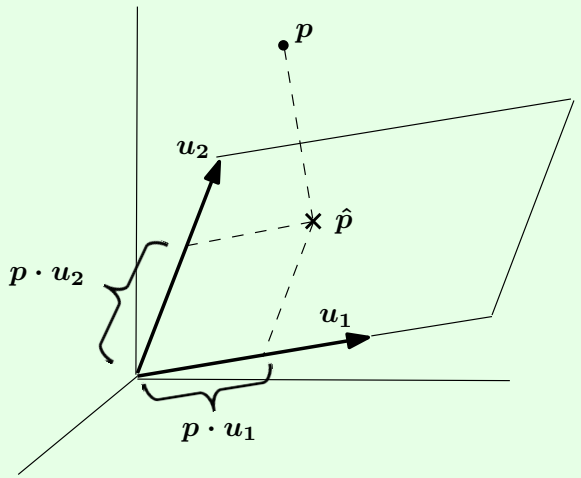


Projection

Let $\boldsymbol{p}$ be a point (a.k.a., a vector) in $\mathbb{R}^d$.
Consider a k-subspace Σ with basis $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_k\}$.
Let $\hat{\boldsymbol{p}}$ be the projection of $\boldsymbol{p}$ onto Σ.

Lemma 1: $|\boldsymbol{p}|^2 = |\boldsymbol{p} - \hat{\boldsymbol{p}}|^2 + \sum_{i=1}^{k} (\boldsymbol{p} \cdot \boldsymbol{u}_i)^2$.

Proof: Follows immediately from $|\hat{\boldsymbol{p}}|^2 = \sum_{i=1}^{k} (\boldsymbol{p} \cdot \boldsymbol{u}_i)^2$ and $|\boldsymbol{p}|^2 = |\boldsymbol{p} - \hat{\boldsymbol{p}}|^2 + |\hat{\boldsymbol{p}}|^2$. □



P : a set of n points in $\mathbb{R}^d$ **whose geometry center is the origin**.

Think: If the geometry center of P is not the origin, what transformations do you need to make this happen?

Goal: Find a k-subspace Σ to minimize its **reconstruction error** defined as

$$\frac{1}{n}\sum_{\boldsymbol{p}\in P}|\boldsymbol{p}-\hat{\boldsymbol{p}}|^2$$

where $\hat{\boldsymbol{p}}$ is the projection of $\boldsymbol{p}$ onto Σ.

Recall that the geometry center of our point set P is the origin.

Define the **covariance** of P between dimensions i and j ($i, j \leq d$) as

$$\frac{1}{n} \sum_{p \in P} p[i] \cdot p[j].$$

where $p[i]$ is the coordinate of p on dimension i.

Think: What is the coordinate mean of the points in P on dimension i?

The **covariance matrix A** of P is a $d \times d$ matrix where $\mathbf{A}[i, j]$ ($i, j \in [1, d]$) is the covariance of P between dimensions i and j.

Note that $\mathbf{A}$ is symmetric, namely, $\mathbf{A} = \mathbf{A}^T$.

Let $\mathbf{A}$ be the covariance matrix defined on the previous slide. If

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

for a $d \times 1$ non-zero vector $\boldsymbol{v}$ and some real value λ, we call

- λ an **eigenvalue** of $\mathbf{A}$, and
- $\boldsymbol{v}$ an **eigenvector** of $\mathbf{A}$.

We also say that $\boldsymbol{v}$ is an **eigenvector for** λ.

Lemma 2: Matrix $\mathbf{A}$ has d eigenvalues $\lambda_1, ..., \lambda_d$. For each $i \in [1, d]$, denote by $\boldsymbol{v}_i$ an eigenvector for λ_i. Then:

- $\lambda_1, ..., \lambda_d$ are all real values.
- For any distinct $i, j \in [1, d]$, the vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$ are orthogonal to each other.

The lemma is a well-known result from linear algebra (proof omitted).

Principle Component Analysis (PCA)

PCA (P, k)

/* The geometry center of P is at the origin */

1. $\mathbf{A} \leftarrow$ the covariance matrix of P
2. compute the eigenvalues of $\mathbf{A}$ and sort them in descending order: $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_d$
3. compute a **unit** eigenvector $\boldsymbol{v}_i$ for λ_i for each $i \in [1, d]$
4. return $\boldsymbol{v}_1, ..., \boldsymbol{v}_k$

The rest of the slides will prove:

Theorem 1: The k-subspace with $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_k\}$ as the basis has the smallest the reconstruction error among all k-subspaces.

Variance along a Direction

Fix an arbitrary unit vector $\boldsymbol{u}$. Define the **variance of $\boldsymbol{u}$** as

$$\frac{1}{n}\sum_{\boldsymbol{p}\in P}|\boldsymbol{p}\cdot\boldsymbol{u}|^2.$$

Indeed, the above is precisely the variance of the set $\{\boldsymbol{p}\cdot\boldsymbol{u} \mid \boldsymbol{p}\in P\}$.

Think: What is the mean of $\{\boldsymbol{p}\cdot\boldsymbol{u} \mid \boldsymbol{p}\in P\}$?

Variance along a Direction

Lemma 3: For any unit vector $\boldsymbol{u}$, the variance of $\boldsymbol{u}$ equals

$$\boldsymbol{u}^T \mathbf{A} \boldsymbol{u}$$

where $\mathbf{A}$ is the covariance matrix of P.

Proof: Define $\mathbf{P}$ as the $n \times d$ matrix whose row $i \in [1, n]$ is $\boldsymbol{p}^T$, where $\boldsymbol{p}$ is the i-th point of P (note: $\boldsymbol{p}$ is a $d \times 1$ vector and hence $\boldsymbol{p}^T$ is a $1 \times d$ vector). Rudimentary calculation shows

$$\begin{aligned} \text{variance of } \boldsymbol{u} &= \frac{1}{n}(\mathbf{P}\boldsymbol{u})^T(\mathbf{P}\boldsymbol{u}) \\ &= \boldsymbol{u}^T \left(\frac{1}{n}\mathbf{P}^T\mathbf{P}\right) \boldsymbol{u} \\ &= \boldsymbol{u}^T \mathbf{A} \boldsymbol{u} \end{aligned}$$

as claimed □

Variance along a Direction

Lemma 4: If λ is an eigenvalue of $\mathbf{A}$ and $\boldsymbol{v}$ is a unit eigenvector for λ, then the variance of $\boldsymbol{v}$ is λ.

Proof:

$$\text{variance of } \boldsymbol{v} = \boldsymbol{v}^T \mathbf{A} \boldsymbol{v} = \boldsymbol{v}^T (\mathbf{A}\boldsymbol{v}) = \lambda \boldsymbol{v}^T \boldsymbol{v} = \lambda.$$

□

Total Variance of a Subspace

Consider a k-subspace Σ with a basis $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_k\}$. We define the **total variance** of Σ as

$$\sum_{i=1}^{k} \text{variance of } \boldsymbol{u}_i.$$

Total Variance vs. Construction Error

Lemma 5: Consider a k-subspace Σ with a basis $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_k\}$. It holds that

$$\frac{1}{n}\sum_{\boldsymbol{p}\in P}|\boldsymbol{p}|^2 = \text{total variance of } \Sigma \ + \ \text{construction error of } \Sigma.$$

Proof: Follows immediately from Lemma 1 and the definitions of total variance (Slide 15) and construction error (Slide 7). □

Implication: We need to maximize total variance to minimize construction error.

The rest of the slides will prove:

Theorem 2: The k-subspace with $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_k\}$ as the basis has the largest total variance among all k-subspaces.

Theorem 1 is then an immediate corollary of Theorem 2 and Lemma 5.

Lemma 6: Among all unit vectors, the vector $\boldsymbol{v}_1$ (the unit eigenvector for the largest eigenvalue λ_1) has the largest variance.

Proof: By Lemma 3, finding a unit vector of the largest variance can be modeled as the following optimization problem:

$$\text{find a vector } \boldsymbol{w} \text{ to maximize } \boldsymbol{w}^T\mathbf{A}\boldsymbol{w} \text{ subject to } \boldsymbol{w}^T\boldsymbol{w} = 1.$$

We can solve the problem using the method of Lagrange multipliers. Introduce a real value λ, and define the Lagrangian function:

$$f(\boldsymbol{w}, \lambda) \quad = \quad \boldsymbol{w}^T\mathbf{A}\boldsymbol{w} - \lambda(\boldsymbol{w}^T\boldsymbol{w} - 1)$$

which yields

$$\frac{\partial f}{\partial \boldsymbol{w}} \quad = \quad 2\mathbf{A}\boldsymbol{w} - 2\lambda\boldsymbol{w}.$$

Setting $\frac{\partial f}{\partial \boldsymbol{w}} = 0$ gives a condition that an optimal $\boldsymbol{w}$ must satisfy:

$$\mathbf{A}\boldsymbol{w} = \lambda \boldsymbol{w}.$$

In other words, λ is an eigenvalue and $\boldsymbol{w}$ is a unit eigenvector of λ.

Lemma 4 shows that the variance of $\boldsymbol{w}$ is exactly λ. Thus, the variance of $\boldsymbol{w}$ cannot exceed the maximum eigenvalue λ_1 of $\mathbf{A}$. Setting $\boldsymbol{w}$ to $\boldsymbol{v}_1$ achieves this variance. □

Lemma 7: Among all the unit vectors orthogoal to $\boldsymbol{v}_1$, the vector $\boldsymbol{v}_2$ (the unit eigenvector for the second largest eigenvalue λ_2) has the largest variance.

Proof: By Lemma 3, the goal is to solve the following optimization problem:

find a vector $\boldsymbol{w}$ to maximize $\boldsymbol{w}^T\mathbf{A}\boldsymbol{w}$ subj. to $\boldsymbol{w}^T\boldsymbol{w} = 1$ and $\boldsymbol{w}^T\boldsymbol{v}_1 = 0$.

Again, we can solve the problem using the method of Lagrange multipliers. Introduce real values λ and ϕ, and define the Lagrangian function:

$$f(\boldsymbol{w}, \lambda, \phi) \quad = \quad \boldsymbol{w}^T\mathbf{A}\boldsymbol{w} - \lambda(\boldsymbol{w}^T\boldsymbol{w} - 1) - \phi\boldsymbol{w}^T\boldsymbol{v}_1$$

which yields

$$\frac{\partial f}{\partial \boldsymbol{w}} \quad = \quad 2\mathbf{A}\boldsymbol{w} - 2\lambda\boldsymbol{w} - \phi\boldsymbol{v}_1.$$

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

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

Next, we argue that ϕ must be 0. The above equation leads to:

$$2\boldsymbol{v}_1^T\mathbf{A}\boldsymbol{w} - 2\lambda\boldsymbol{v}_1^T\boldsymbol{w} - \phi\boldsymbol{v}_1^T\boldsymbol{v}_1 \quad = \quad 0. \tag{2}$$

Recall that $\boldsymbol{v}_1^T\boldsymbol{w} = 0$ and $\boldsymbol{v}_1^T\boldsymbol{v}_1 = 1$. Furthermore:

$$\begin{aligned} \boldsymbol{v}_1^T\mathbf{A}\boldsymbol{w} &= \boldsymbol{w}^T\mathbf{A}^T\boldsymbol{v}_1 \quad (\text{transposing a } 1 \times 1 \text{ matrix}) \\ &= \boldsymbol{w}^T\mathbf{A}\boldsymbol{v}_1 = \boldsymbol{w}^T(\mathbf{A}\boldsymbol{v}_1) = \lambda_1\boldsymbol{w}^T\boldsymbol{v}_1 = 0. \end{aligned}$$

Hence, from (2), we get $\phi = 0$. Thus, (1) now becomes:

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

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

It thus follows from Lemma 4 that the optimal $\boldsymbol{w}$ must be $\boldsymbol{v}_2$. □

The following theorem generalizes Lemma 7.

Theorem 3: For any $t \in [2, d]$, the vector $\boldsymbol{v}_t$ (the unit eigenvector for the t-th largest eigenvalue λ_t) has the largest variance among all the unit vectors orthogonal to $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{t-1}$.

The proof is a straightforward extension of the proof of Lemma 7 and is left to you as an exercise.

We are now ready to prove Theorem 2 using mathematical induction.

Base case. For $k = 1$, the theorem's correctness follows directly from Lemma 6.

Inductive case. Assuming the the theorem's correctness on $k = t - 1$ for some $t \in [2, d]$, we will show that it also holds on $k = t$.

Let Σ^* be the t-subspace that has $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_t\}$ as a basis.
Let Σ be an arbitrary t-subspace.

Our goal is to show that the total variance of Σ is at most that of Σ^*.

Lemma 8: We can find a vector $\boldsymbol{w}$ in the t-subspace Σ such that $\boldsymbol{w}$ is orthogoal to $\boldsymbol{v}_i$ for every $i \in [1, t-1]$.

Proof: Let $\{\boldsymbol{u}_1, ..., \boldsymbol{u}_t\}$ be an arbitrary basis of Σ. Any vector $\boldsymbol{w}$ in Σ is in the span of $\boldsymbol{u}_1, ..., \boldsymbol{u}_t$. This means

$$\boldsymbol{w} = \Sigma_{i=1}^{t} c_i \boldsymbol{u}_i$$

for some real values $c_1, c_2, ..., c_k$. As $\boldsymbol{w}$ needs to be orthogonal to each of $\boldsymbol{v}_1, ..., \boldsymbol{v}_{t-1}$, we have:

$$\begin{aligned} \boldsymbol{w} \cdot \boldsymbol{v}_1 &= 0 \\ \boldsymbol{w} \cdot \boldsymbol{v}_2 &= 0 \\ &... \\ \boldsymbol{w} \cdot \boldsymbol{v}_{t-1} &= 0 \end{aligned}$$

This yields a set of $t-1$ linear equations for the t variables $c_1, ..., c_t$. The linear system has infinitely many solutions. □

Lemma 8 implies that Σ has a basis $\{\boldsymbol{w}_1, ..., \boldsymbol{w}_t\}$ such that $\boldsymbol{w}_t$ is orthogonal to all of $\boldsymbol{v}_1, ..., \boldsymbol{v}_{t-1}$ (**think: why?**). We can then assert by Theorem 3 that

$$\text{variance of } \boldsymbol{v}_t \geq \text{variance of } \boldsymbol{w}_t.$$

Let Σ^*_{t-1} be the subspace defined by the basis $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_{t-1}\}$.
Let Σ_{t-1} be the subspace defined by the basis $\{\boldsymbol{w}_1, ..., \boldsymbol{w}_{t-1}\}$.

By the inductive assumption, we have

$$\text{total variance of } \Sigma^*_{t-1} \geq \text{total variance of } \Sigma_{t-1}.$$

Therefore

$$\begin{aligned}
\text{total variance of } \Sigma^* &= \text{total variance of } \Sigma^*_{t-1} + \text{variance of } \boldsymbol{v}_t \\
&\geq \text{total variance of } \Sigma_{t-1} + \text{variance of } \boldsymbol{v}_t \\
&\geq \text{total variance of } \Sigma_{t-1} + \text{variance of } \boldsymbol{w}_t \\
&= \text{total variance of } \Sigma.
\end{aligned}$$

This completes the proof of Theorem 2.