Finding Principal Directions

Mathematics of Machine Learning

Math 3180 — University of Connecticut

The Data Matrix

- X is an $N \times k$ matrix:
  - rows $\longleftrightarrow$ samples
  - columns $\longleftrightarrow$ features
- **Assumption:** the data is *centered* — each column sums to zero.
- Each row of X is a point in $\mathbb{R}^k$.
  The N rows form a *cloud of points* in $\mathbb{R}^k$.

Variance in a Direction

- Let $u \in \mathbb{R}^k$ be a **unit vector**, $\|u\|^2 = 1$.

- The *variance of the data in direction u* is

$$\sigma_u^2 \;=\; u^T D_0\, u,$$

where $D_0 = \dfrac{1}{N} X^T X$ is the **covariance matrix**.

- $\sigma_u^2 = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} (x_i \cdot u)^2$ is the mean squared projection onto u.

u is a unit vector in $\mathbb{R}^k$

$\sigma_u^2 = u^T D_0 u$

$D_0 = \frac{1}{N} X^T X$ covariance matrix.

$$\sigma_u^2 = [u_1, u_2] \begin{bmatrix} d_{00} & d_{01} \\ d_{10} & d_{11} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$[d_{00}u_1 + d_{10}u_2 \quad d_{01}u_1 + d_{11}u_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$$

$$d_{00}u_1^2 + d_{10}u_1u_2 + d_{01}u_1u_2 + d_{11}u_2^2$$

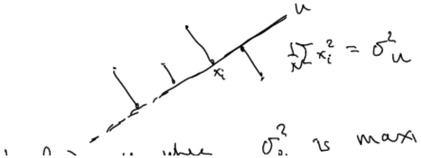
$$\sigma_u^2 = d_{00}u_1^2 + 2d_{10}u_1u_2 + d_{11}u_2^2$$

$$Au_1^2 + Bu_1u_2 + Cu_2^2 = 1$$

The Optimization Problem

Goal

Find the unit vector $u \in \mathbb{R}^k$ that **maximizes** $\sigma_u^2 = u^T D_0\, u$.

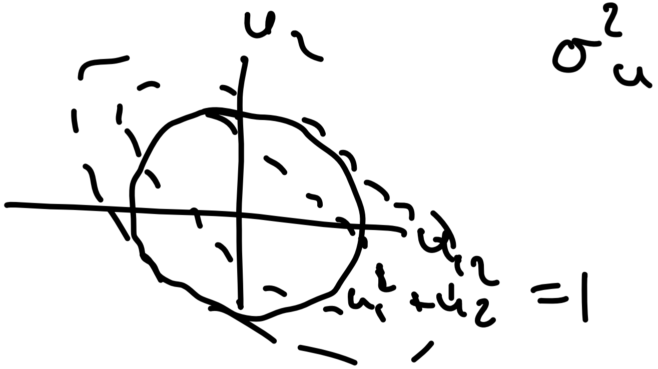


Projections $x_i \cdot u$ of data points onto u; their mean square is σ_u^2.

- This is a *constrained* optimization problem:

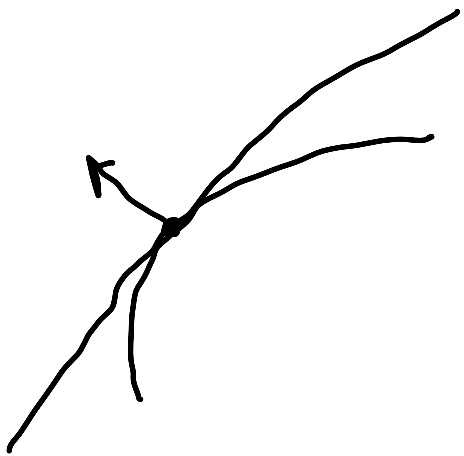
$$\text{maximize } u^T D_0 u \quad \text{s.t.} \quad \|u\|^2 = 1.$$

- Tool: **Lagrange multipliers**.



Lagrange Multipliers: General Setup

- **Objective**: $F(x_1, \ldots, x_k)$ — the function to maximize.
- **Constraint**: $g(x_1, \ldots, x_k) = 0$.



Form the **Lagrangian**:

$$S(x_1, \ldots, x_k, \lambda) \;=\; F - \lambda\, g.$$

Critical points satisfy:

$$\frac{\partial S}{\partial x_i} = 0 \quad (i = 1, \ldots, k) \qquad \text{and} \qquad \frac{\partial S}{\partial \lambda} = 0 \iff g = 0.$$

$$\nabla S = 0 \iff \nabla F = \lambda \nabla g$$

Setting Up the Lagrangian

Write $D_0 = (d_{ij})$. Then

$$\sigma_u^2 \;=\; u^T D_0 u \;=\; \sum_{r=1}^{k} \sum_{s=1}^{k} u_r \, d_{rs} \, u_s .$$

The constraint is $g(u) = u_1^2 + \cdots + u_k^2 - 1 = 0$.

The Lagrangian is:

$$S(u, \lambda) \;=\; \sum_{r=1}^{k} \sum_{s=1}^{k} u_r \, d_{rs} \, u_s \;-\; \lambda\big(u_1^2 + \cdots + u_k^2 - 1\big) .$$

Computing the Partial Derivatives

Differentiating S with respect to u_i and using the symmetry $d_{rs} = d_{sr}$ of D_0:

$$\frac{\partial S}{\partial u_i} \;=\; 2\sum_{s=1}^{k} d_{is}\, u_s \;-\; 2\lambda u_i.$$

In vector form:

$$\nabla_u S \;=\; 2\, D_0\, u \;-\; 2\lambda\, u \;=\; 2\,(D_0 - \lambda I)\, u.$$

The condition $\dfrac{\partial S}{\partial \lambda} = 0$ gives $\|u\|^2 = 1$.

The Critical Point Equations

Setting $\nabla_u S = 0$ and $\partial S / \partial \lambda = 0$ gives the system:

$$\begin{cases} (D_0 - \lambda I)\, u = 0, \\ \|u\|^2 = 1. \end{cases}$$

The first equation is equivalent to

$$\boxed{D_0\, u = \lambda\, u.}$$

u is an eigenvector of D_0 with eigenvalue λ, normalized so that $\|u\| = 1$.

The Value of σ_u^2 at a Critical Point

If $D_0 u = \lambda u$ and $\|u\| = 1$, then

$$\sigma_u^2 \;=\; u^T D_0 u \;=\; u^T(\lambda u) \;=\; \lambda \, \|u\|^2 \;=\; \lambda.$$

The variance in direction u *equals* the corresponding eigenvalue.

Main Theorem

Theorem

The critical points of $\sigma_u^2 = u^T D_0 u$ subject to $\|u\|^2 = 1$ are exactly the unit eigenvectors of D_0. At such a critical point, $\sigma_u^2 = \lambda$ (the corresponding eigenvalue).

- The **maximum variance** equals the **largest eigenvalue** of D_0; it is achieved along the corresponding eigenvector.
- The **minimum variance** equals the **smallest eigenvalue** of D_0.

Fact

$D_0 = \frac{1}{N} X^T X$ is symmetric, so all its eigenvalues are real (and in fact non-negative).