Principal Component Analysis

Suppose we take n measurements of m variables (with real values).
Each measurement is a vector in $\mathbb{R}^m$, and our n measurements are then n vectors $x_1, \dots, x_n \in \mathbb{R}^m$

$\{$point cloud$\}$

The mean (or average) of these vectors is

$$\mu := \frac{1}{n}(x_1 + \cdots + x_n).$$

Note The i-th component of μ is just the average value of the i-th variable: $\mu_i = \frac{1}{n}(x_{1i} + \cdots + x_{ni})$.

Q How can we mimic other important statistics?

e.g. For $a_1, \dots, a_n \in \mathbb{R}$, their variance is

$$\mathrm{var}(a) = \frac{1}{n-1}\left((a_1-\mu)^2 + \cdots + (a_n-\mu)^2\right).$$

If we also measure $b_1, \dots, b_n \in \mathbb{R}$, the covariance b/w a_i, b_i is

$$\mathrm{cov}(a,b) = \frac{1}{n-1}\left((a_1-\mu_a)(b_1-\mu_b) + \cdots + (a_n-\mu_a)(b_n-\mu_b)\right).$$

- Variance measures how much the a_i differ from their mean and its square root is the standard deviation.
- Covariance measures how the a_i & b_i depend on each other. E.g. negative cov arises when large a_i predicts small b_i (relative to means).

Going back to multivariate setting, define

$$B = \begin{pmatrix} x_1-\mu & x_2-\mu & \cdots & x_n-\mu \end{pmatrix}$$

the $m \times n$ matrix with i-th column $x_i - \mu$. This is the recentering of the data so the mean is 0.

Defn The covariance matrix $S = \frac{1}{n-1} BB^T$.

Note $S \in M_{m\times m}(\mathbb{R})$ is symmetric.

e.g. $x_1 = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{pmatrix}$ $x_2 = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}$ $x_3 = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{pmatrix}$ $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{pmatrix}$

$$B = \begin{pmatrix} a_1-\mu_1 & b_1-\mu_1 & c_1-\mu_1 \\ a_2-\mu_2 & b_2-\mu_2 & c_2-\mu_2 \\ a_3-\mu_3 & b_3-\mu_3 & c_3-\mu_3 \\ a_4-\mu_4 & b_4-\mu_4 & c_4-\mu_4 \end{pmatrix}$$

Then $S_{11} = \frac{1}{3-1}\left((a_1-\mu_1)^2 + (b_1-\mu_1)^2 + (c_1-\mu_1)^2\right)$ = variance of first variable.

Similarly, S_{ii} = variance of i-th variable.

Also $S_{21} = \frac{1}{3-1}\left((a_1-\mu_1)(a_2-\mu_2) + (b_1-\mu_1)(b_2-\mu_2) + (c_1-\mu_1)(c_2-\mu_2)\right)$

= covariance of first and second variables.

Similarly, S_{ij} = cov of i-th & j-th vars.

Defn The total variance is $tr(S) = \Sigma$ var of variables.

e.g. $m=2$

Observe: S_{11} large, S_{22} small

covariance small

$S = \begin{pmatrix} 95 & 1 \\ 1 & 5 \end{pmatrix}$ (total var 100)

$S = \begin{pmatrix} 50 & 40 \\ 40 & 50 \end{pmatrix}$

Goal Recognize the ~~sim~~ similarity of these data sets with linear algebra.

Spectral Thm: If $A \in M_{n\times n}(\mathbb{R})$ and $A = A^T$, then A is orthogonally diagonalizable with real eigenvalues. I.e. $\exists \lambda_1, \dots, \lambda_n \in \mathbb{R}$ and orthogonal nonzero vectors $v_1, \dots, v_n \in \mathbb{R}^n$ s.t. $Av_i = \lambda_i v_i$.

Pf: Later.

Note: For $B \in M_{m\times n}(\mathbb{R})$, BB^T and B^TB are symmetric real matrices to which the spectral thm applies.

Prop: BB^T and B^TB share the same nonzero eigenvalues.

Pf: Take v an eigenvector of B^TB with eigenvalue $\lambda \neq 0$, so that

$$B^TBv = \lambda v.$$

Mult on left by B to get

$$BB^T(Bv) = \lambda (Bv).$$

Hence λ is an eigenvalue of BB^T with eigenvector Bv.
(Indeed, $Bv \neq 0$ since $B^TBv = \lambda v \neq 0$.)

Similarly $BB^Tw = \lambda w \implies B^TB(B^Tw) = \lambda(B^Tw)$

so BB^T, B^TB have the same nonzero eigenvalues. $\square$

TPS: What if B is 500×2?

(Find eigenvalues of 2×2 matrix B^TB. These are eigenvalues of BB^T (a 500×500 matrix) and all others are 0.)

Prop: The eigenvalues of BB^T and B^TB are all nonnegative.

Pf: Take v an eigenvector of B^TB with eigenvalue λ. Then

$$\begin{aligned}\|Bv\|^2 &= (Bv)\cdot(Bv) = (Bv)^T(Bv) \\ &= v^T(B^TB)v \\ &= v^T(\lambda v) \\ &= \lambda v^Tv \\ &= \lambda \|v\|^2.\end{aligned}$$

Since $\|Bv\|^2 \geq 0$ and $\|v\|^2 \neq 0$, must have $\lambda \geq 0$.

$\square$