Random Variable

R.V. is a function that maps random event to a number. It is not a variable in the sense that is appear in the equations.

Ex 1:

$X : \{\text{it will rain, it will not rain}\} \rightarrow \{1, 0\}$

Ex 2:

$X : \{\text{Head, Tail}\} \rightarrow \{1, 0\}$

Ex 3:

$X : \{\text{Rolling a dice}\} \rightarrow \{\text{Number that faces up}\}$

R.V.

- **Discrete**: Previous three examples
- **Continuous**: X = Exact amount of rain in m^2

Expected Value

$$E[X = x] = \sum_x x\,p(x) = \mu$$

Expectation of other functions can be calculated

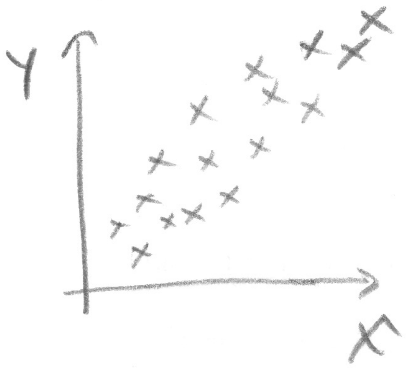
$$E[g(X)] = \sum_x g(x)\,p(x)$$

Variance:

$$E[(X-\mu)^2] = \sum_x (x-\mu)^2 \, p(x) = Var(X)$$
$$= \sigma^2$$

Covariance

- Positive covariance between two r.v.s

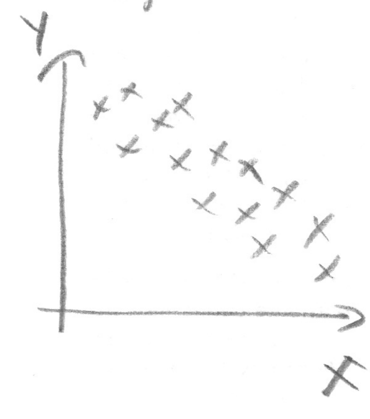


$$Cov(X,Y) = E[(X-\mu_x)(Y-\mu_y)]$$

$X > \mu_x$: $+$ $\quad + \quad$ | $+$

$X < \mu_x$: $-$ $\quad - \quad$ | $+$

- Negative covariance between two r.v.s



$X > \mu_x$: $+$ $\quad - \quad$: $-$

$X < \mu_x$: $-$ $\quad + \quad$: $-$

PCA

PCA tries to find "components" that capture the maximal variance within the data.

Ex: $X \to I \times J$ matrice that contains our I observations with J features.

If we subtract mean of each column from the corresponding column, the data points are centered

$$\vec{x}_i \leftarrow \vec{x}_i - \vec{\mu}_x$$ → subtract mean from each row

So, we assume that in the following

X is our centered dataset

* Finding direction of maximal variance corresponds to

$$\max_{\vec{c}} \vec{c}^T X^T X \vec{c}, \quad c^T c = 1$$

- If we think each observation x_i, $i = 1 \ldots I$ separately
↳ This is row vector.

$$w = x_i \vec{c}$$

Projects observation onto c. This is a scalar on the direction c.

Therefore $\vec{w} = X\vec{c}$ → projects dataset onto c

Because X contains mean subtracted data

$\vec{w}^T \vec{w}$ → corresponds to variance of the projected data

It is equal to $\vec{c}^T X^T X \vec{c} = \vec{w}^T \vec{w}$

Output of the $\vec{c}^T X^T X \vec{c}$ is a scaler which corresponds to variance in the direction of $\vec{c}$.

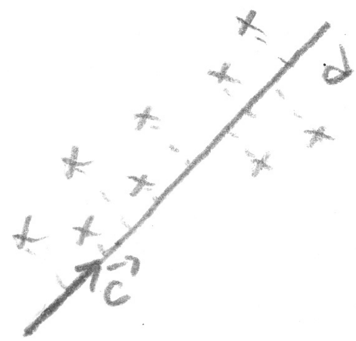
To find maximum of the problem

$$\vec{c}^T X^T X \vec{c}$$

we will compute the max eigenvalue of $X^T X$ and set $\vec{c}$ to the eigenvector of $X^T X$

* why do we have to find greatest eigenvalue?

Projecting data on a vector $\vec{c}$ means that projecting data on a line in the orientation of the vector $\vec{c}$



The line d is parametricaly represented as

$$\underline{d = \vec{m} + a\vec{c}} \quad , \quad c^T c = 1 \rightarrow \text{unit vector}$$

↳ parametric equation of a line

The points projected on this straight line would be

$$\vec{x}_i = \vec{m} + a_i c, \quad i = 1 \ldots I$$

- To minimize error we have to minimize following equation

$$J_1(a_1, a_2, \dots, a_N, \vec{e}) = \sum_{i=1}^{N} \| \vec{m} + a_i \vec{e} - \vec{x}_i \|^2$$

Expand:

$$J_1 = \sum_{i=1}^{N} a_i^2 \|\vec{e}\|^2 - 2\sum_{i=1}^{N} a_i \vec{e}\,(\vec{x}_i - \vec{m}) + \sum_{i=1}^{N} \|\vec{x}_i - \vec{m}\|^2$$

$$= \sum_{i=1}^{N} a_i^2 - 2\sum_{i=1}^{N} a_i \vec{e}^T (\vec{x}_i - \vec{m}) + \sum_{i=1}^{N} \|\vec{x}_i - \vec{m}\|^2$$

It is the quadratic equation w.r.t. a_i

$$\frac{\partial J_1}{\partial a_i} = a_i - \vec{e}^T (\vec{x}_i - \vec{m}) = 0$$

$$a_i = \vec{e}^T (\vec{x}_i - \vec{m})$$

$$J_1(\vec{e}) = -\sum \vec{e}^T (\vec{x}_i - \vec{m})(\vec{x}_i - \vec{m}) \vec{e} + \sum_{i=1}^{N} \|\vec{x}_i - \vec{m}\|^2$$

$$= -\vec{e}^T S \vec{e} + \sum_{i=1}^{N} \|\vec{x}_i - \vec{m}\|^2$$

where

$$S = \sum_{i=1}^{N} (\vec{x}_i - \vec{m})(x_i - m)^T$$ → called sample covariance matrix

So, minimizing J_1 is equal to maximize

$\vec{c}^T S \vec{c}$, subject to $c^T c = 1$

If we add constraint as a lagrange multiplier

$$\max_{c} \; c^T S c - \lambda (c^T c - 1)$$

quadratic equation

If we differentiate and equate to 0

$$2S\vec{c} - 2\lambda c = 0 \Rightarrow \underbrace{S\vec{c} = \lambda \vec{c}}$$

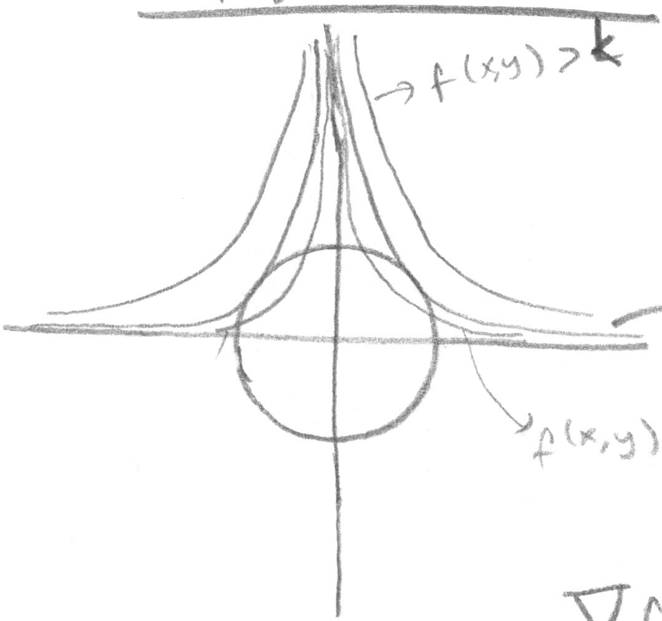
This is the eigenvalue eigenvector problem.

Because $\vec{c}$ is a unit vector

$c^T S c \rightarrow$ scalar value

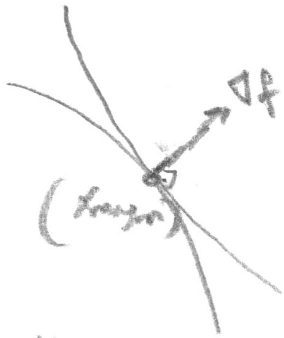
maximum of $c^T S c$ this value corresponds to maximum of eigenvalue λ.

Lagrange Multipliers



Maximize $f(x,y) = x^2y$

on the set $x^2+y^2=1$



∇f (Gradient of f)

gives the direction

the most rapid changes of the

contours.

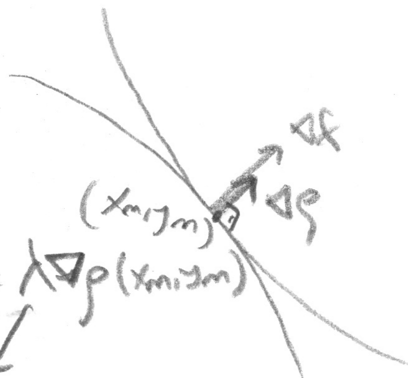
∇f is perpendicular to the contour

lines

Define a constraint as a curve

$g(x,y) = x^2+y^2$

$$\nabla g = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$



$\nabla f(x_m, y_m) = \lambda \nabla g(x_m, y_m)$

λ: Lagrange multiplier

Therefore

$$\nabla f = \begin{bmatrix} 2xy \\ x^2 \end{bmatrix} = \lambda \nabla g = \lambda \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\boxed{\begin{aligned} 2xy &= 2x\lambda \\ x^2 &= 2y\lambda \\ x^2+y^2 &= 1 \end{aligned}}$$

Three variables (unknown)
Three equations

$y = \lambda$ $\qquad$ $3\lambda^2 = 1$

$x = \lambda\sqrt{2}$ $\qquad$ $\lambda = \pm\sqrt{1/3}$

$x = \pm\sqrt{2/3}$

There are four points

$\left(\frac{\sqrt{2}}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$

$\left(-\frac{\sqrt{2}}{\sqrt{3}}, \frac{1}{\sqrt{3}}\right)$

Gives max for $f(x, y)$

$\cancel{\left(\sqrt{2/3}, -\sqrt{1/3}\right)}$

$\cancel{\left(-\sqrt{2}/3, -1/\sqrt{3}\right)}$

Legrangian: Combine constraint function with the target function.

Ex: max $f(x,y) = x^2 e^y y$ subject to $g(x,y) = x^2+y^2 = b$

$\nabla f = \lambda \nabla g$

$$\mathcal{L}(x,y,\lambda) = f(x,y) - \lambda\left(\rho(x,y) - b\right)$$

$$\nabla \mathcal{L} = 0 \rightarrow \nabla \mathcal{L} = \begin{bmatrix} \partial \mathcal{L}/\partial x \\ \partial \mathcal{L}/\partial y \\ \partial \mathcal{L}/\partial \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

$$\frac{\partial \mathcal{L}}{\partial x} = \frac{\partial f}{\partial x} - \lambda \frac{\partial \rho}{\partial x} = 0$$

$$\frac{\partial f}{\partial x} = \frac{\lambda \rho}{\partial x}$$

$$\frac{\partial \mathcal{L}}{\partial y} = \frac{\partial f}{\partial y} - \lambda \frac{\partial \rho}{\partial y} = 0$$

$$\frac{\partial f}{\partial y} = \lambda \frac{\partial \rho}{\partial y}$$

$$\nabla f = \lambda \nabla \rho$$

Lagrangian converts constrained problem into unconstrained problem that can be soluable using computer programs more easily.

$$\frac{\partial \mathcal{L}}{\partial \lambda} = -\left(\rho(x,y) - b\right) = 0$$

$$\rho(x,y) = b$$