Fundamentals of Statistics

Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-635

9/6/2023

Statistics

Standard deviation - a measure of data points differ from mean

Average of differences (w/ mean as reference point)

Higher standard deviation indicates higher spread, less consistency, and less “clustering/blobbing”

Sample standard deviation: $s = \sqrt{\frac{\sum\left(x - \bar{X}\right)^2}{n-1}}$

Population standard deviation: $\sigma = \sqrt{\frac{\sum\left(x - \mu\right)^2}{N}}$

Expectation

X *is univariate here!* →

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

Definition A *random vector* $\vec{X}$ is a vector $(X_1, X_2, \ldots, X_p)$ of jointly distributed random variables. As is customary in linear algebra, we will write vectors as column matrices whenever convenient.

Definition The expectation $E\vec{X}$ of a random vector $\vec{X} = [X_1, X_2, \ldots, X_p]^T$ is given by

$$E\vec{X} = \begin{bmatrix} EX_1 \\ EX_2 \\ \vdots \\ EX_p \end{bmatrix}.$$

$$\begin{aligned} E\vec{X} &= E\begin{bmatrix} X_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + E\begin{bmatrix} 0 \\ X_2 \\ \vdots \\ 0 \end{bmatrix} + \cdots + E\begin{bmatrix} 0 \\ 0 \\ \vdots \\ X_p \end{bmatrix} \\ &= \begin{bmatrix} EX_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ EX_2 \\ \vdots \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ EX_p \end{bmatrix} \\ &= \begin{bmatrix} EX_1 \\ EX_2 \\ \vdots \\ EX_p \end{bmatrix}. \end{aligned}$$

The linearity properties of the expectation can be expressed compactly by stating that for any $k \times p$-matrix A and any $1 \times j$-matrix B,

$$E(A\vec{X}) = AE\vec{X} \quad \text{and} \quad E(\vec{X}B) = (E\vec{X})B.$$

X is *p x 1* here! ***X*** is *j x 1* here!

Variance-Covariance

Definition The *variance–covariance matrix* (or simply the *covariance matrix*) of a random vector $\vec{X}$ is given by:

$$\text{Cov}(\vec{X}) = E\left[(\vec{X} - E\vec{X})(\vec{X} - E\vec{X})^T\right].$$

Proposition

$$\text{Cov}(\vec{X}) = E[\vec{X}\vec{X}^T] - E\vec{X}(E\vec{X})^T.$$

Proposition

$$\text{Cov}(\vec{X}) = \begin{bmatrix} \text{Var}(X_1) & \text{Cov}(X_1, X_2) & \cdots & \text{Cov}(X_1, X_p) \\ \text{Cov}(X_2, X_1) & \text{Var}(X_2) & \cdots & \text{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(X_p, X_1) & \text{Cov}(X_p, X_2) & \cdots & \text{Var}(X_p) \end{bmatrix}.$$

Thus, $\text{Cov}(\vec{X})$ *is a symmetric matrix, since* $\text{Cov}(X, Y) = \text{Cov}(Y, X)$.

Covariance

- Variance and Covariance:
  - Measure of the "spread" of a set of points around their center of mass (mean)
- Variance:
  - Measure of deviation from mean for points in one dimension
- Covariance:
  - Measure of how much each of dimensions vary from mean with **respect to each other**

- **Covariance is measured between two dimensions**
- **Covariance → relation between two dimensions**
- **Covariance between one dimension is variance**

The Gaussian Distribution

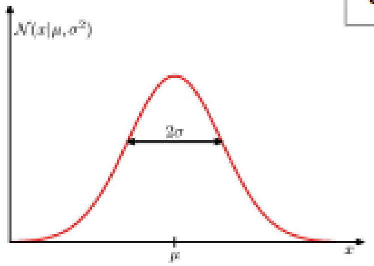
Carl Friedrich Gauss
1777-1855

- For single real-valued variable x

$$N(x \mid \mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

68% of data lies within σ of mean
95% within 2σ

- Parameters:
  - Mean μ, variance σ^2,



- *Standard deviation* σ
- *Precision* $\beta = 1/\sigma^2$, $E[x] = \mu$, $Var[x] = \sigma^2$
- For D-dimensional vector x, multivariate Gaussian

$$N(\mathrm{x} \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathrm{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathrm{x}-\boldsymbol{\mu})\right\}$$

μ is a mean vector, Σ is a D x D covariance matrix, $|\Sigma|$ is the determinant of Σ

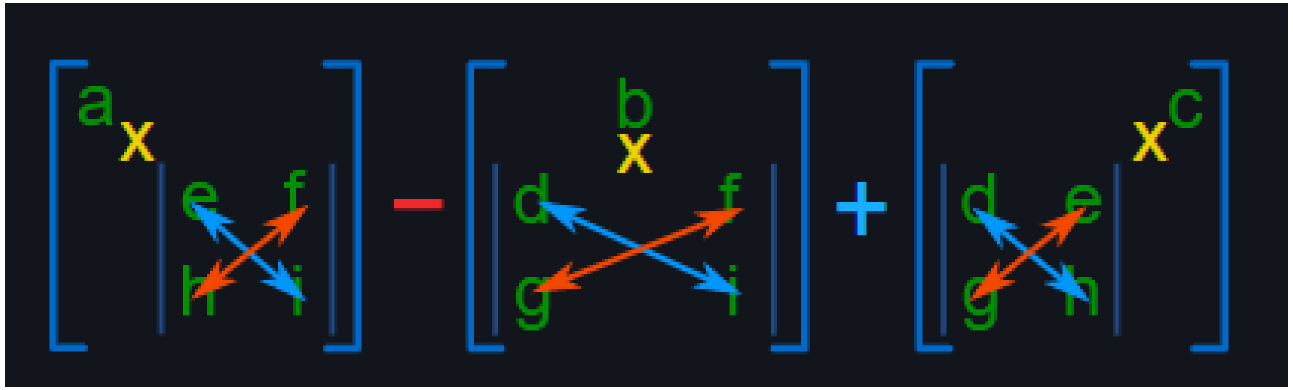
Σ^{-1} is also referred to as the precision matrix

Matrix Determinant

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

$$|A| = a(ei - fh) - b(di - fg) + c(dh - eg)$$

"The determinant of A equals ... etc"

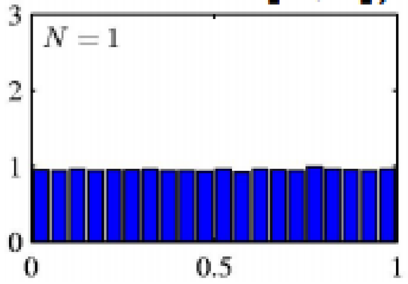


Importance of Gaussian

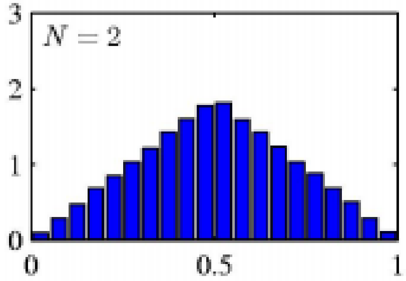
- Gaussian arises in many different contexts, e.g.,
  - Sum of set of random variables becomes increasingly Gaussian

The two values
could be 0.8 and 0.2
whose average is 0.5
More ways of getting 0.5
than say 0.1

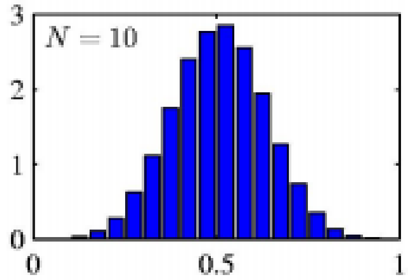
One variable histogram
(uniform over [0,1])



Mean of two variables



Mean of ten variables



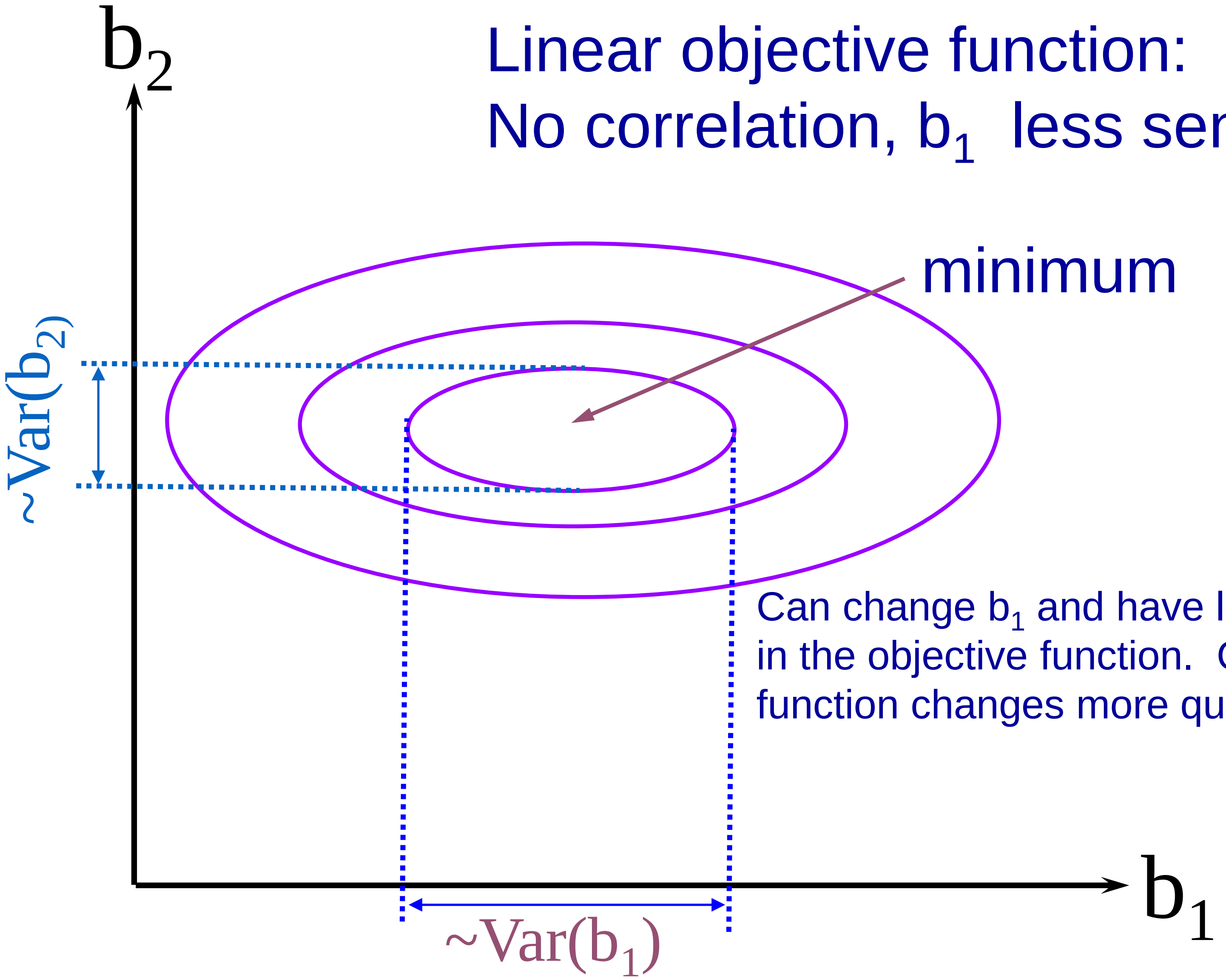
The Central Limit Theorem

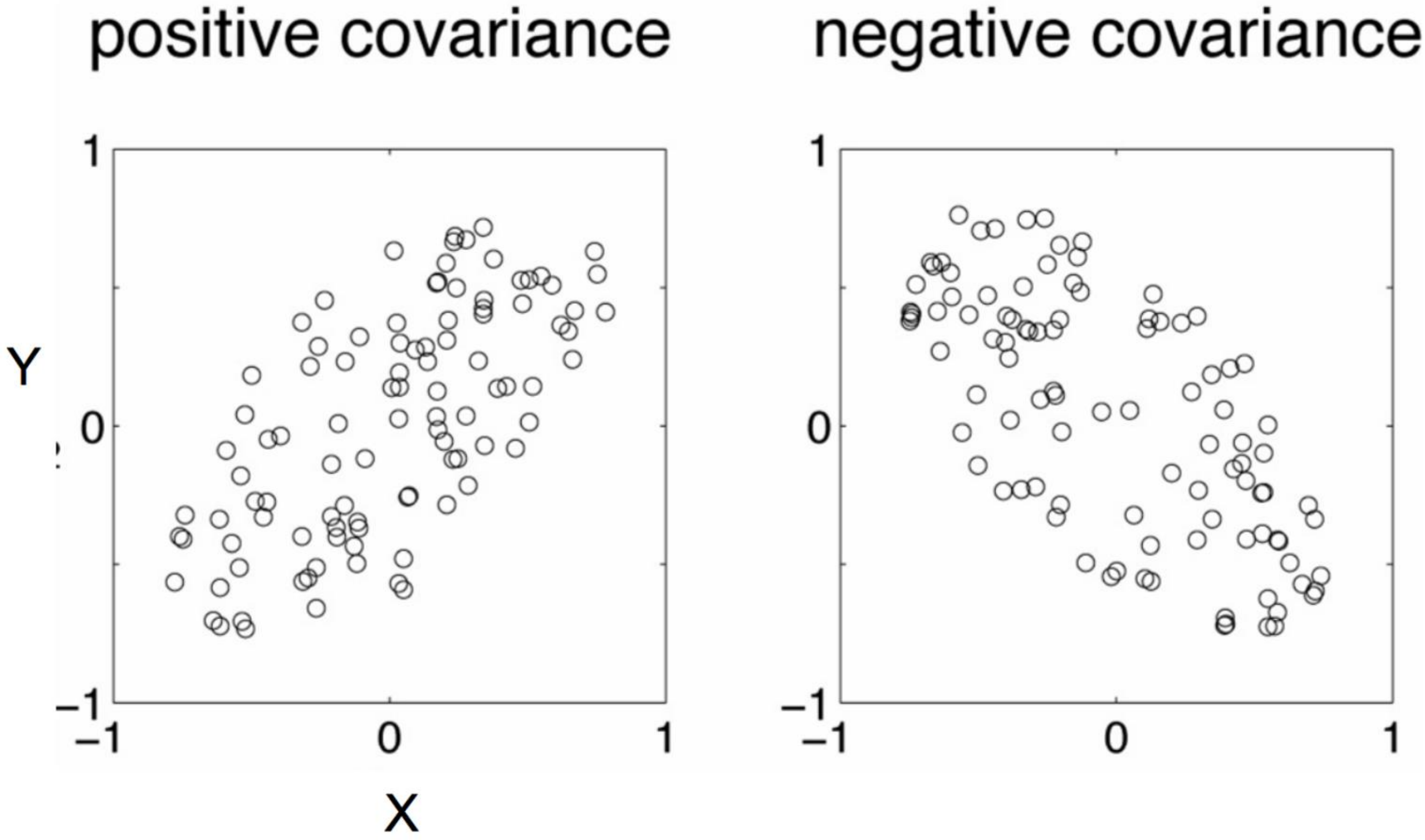
- In many cases, for i.i.d. random variables, sampling distribution of standardized sample mean tends towards a standard normal distribution even if original variables are not normally distributed

Lindeberg–Lévy CLT — Suppose $\{X_1, \ldots, X_n\}$ is a sequence of i.i.d. random variables w/ $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$. Then, as n approaches infinity, the random variables $\sqrt{n}(\bar{X}_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$:

$$\sqrt{n}\left(\bar{X}_n - \mu\right) \xrightarrow{a} \mathcal{N}\left(0, \sigma^2\right).$$

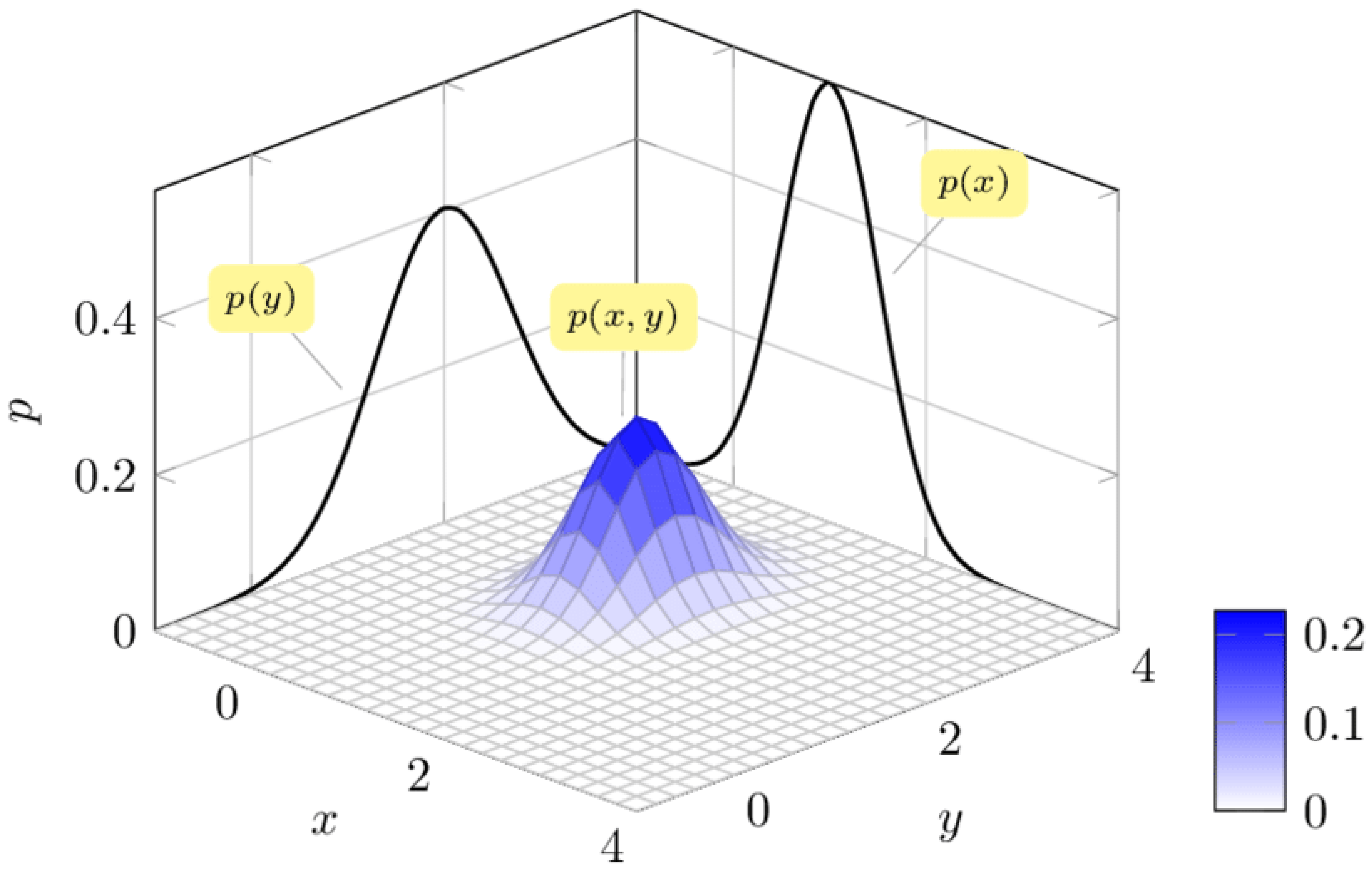
Example: Parameter Variance & Covariance





Positive: Both dimensions increase or decrease together

Negative: While one increase the other decrease



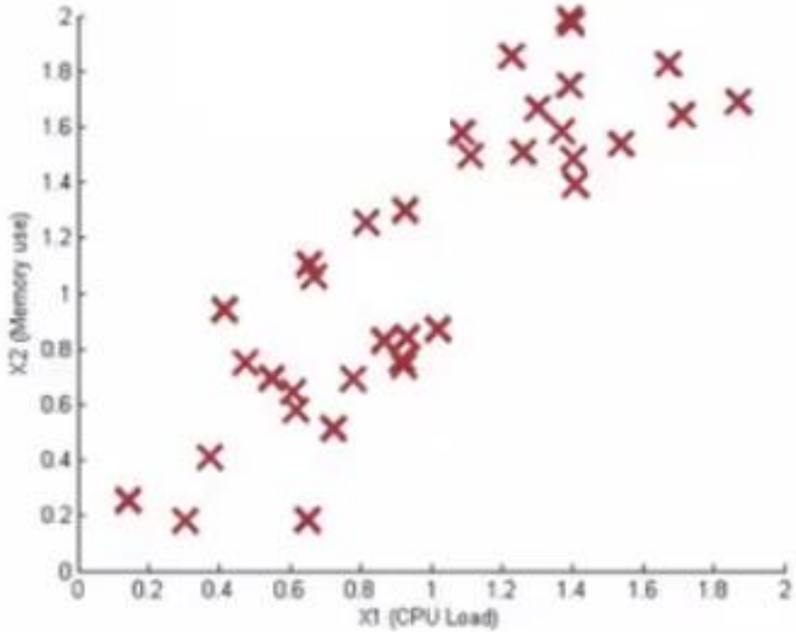
p(y)
p(x, y)
p(x)
p
0.4
0.2
0
0
2
4
x
0
2
4
y
0.2
0.1
0

Anomaly detection with the multivariate Gaussian

1. Fit model $p(x)$ by setting

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$$

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m}(x^{(i)} - \mu)(x^{(i)} - \mu)^T$$



2. Given a new example x, compute

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

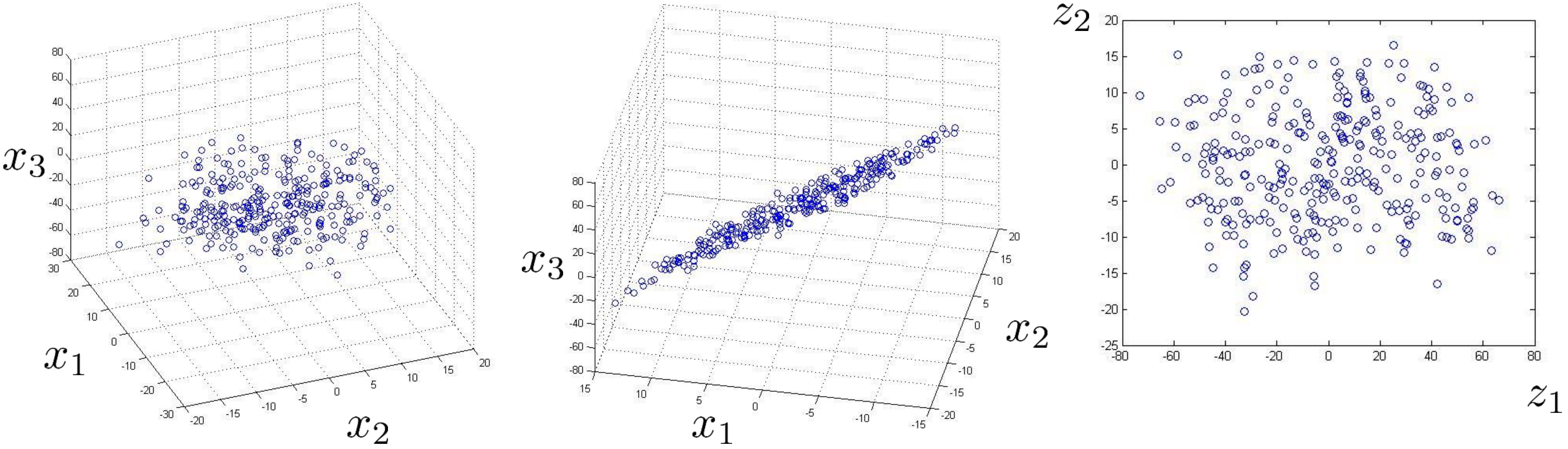
Flag an anomaly if $p(x) < \varepsilon$

Why? Well, Dimensionality Reduction…

- PCA (Principal Component Analysis):
  - Find projection that maximize the variance
- ICA (Independent Component Analysis):
  - Similar to PCA except assumes non-Gaussian features
- Multidimensional Scaling:
  - Find projection that best preserves inter-point distances
- LDA(Linear Discriminant Analysis):
  - Maximizing the component axes for class-separation
- …

Data Compression

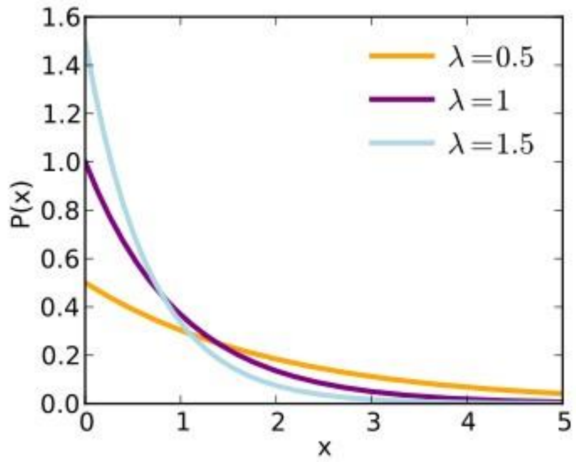
Reduce data from 3D to 2D



More Distributions

Exponential:

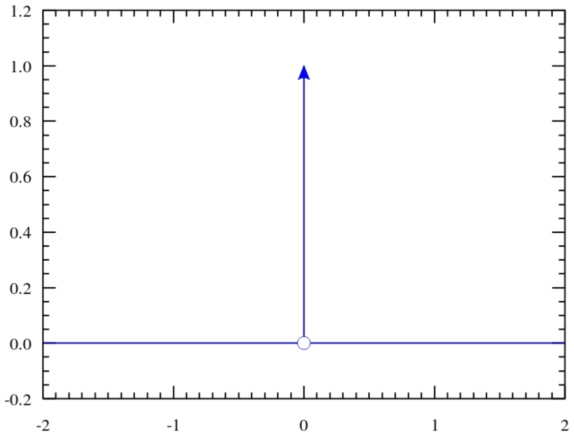
$$p(x;\lambda) = \lambda \mathbf{1}_{x\geq 0} \exp(-\lambda x) .$$



Used to predict the waiting time until the next event occurs, such as a success, failure, or arrival

Dirac:

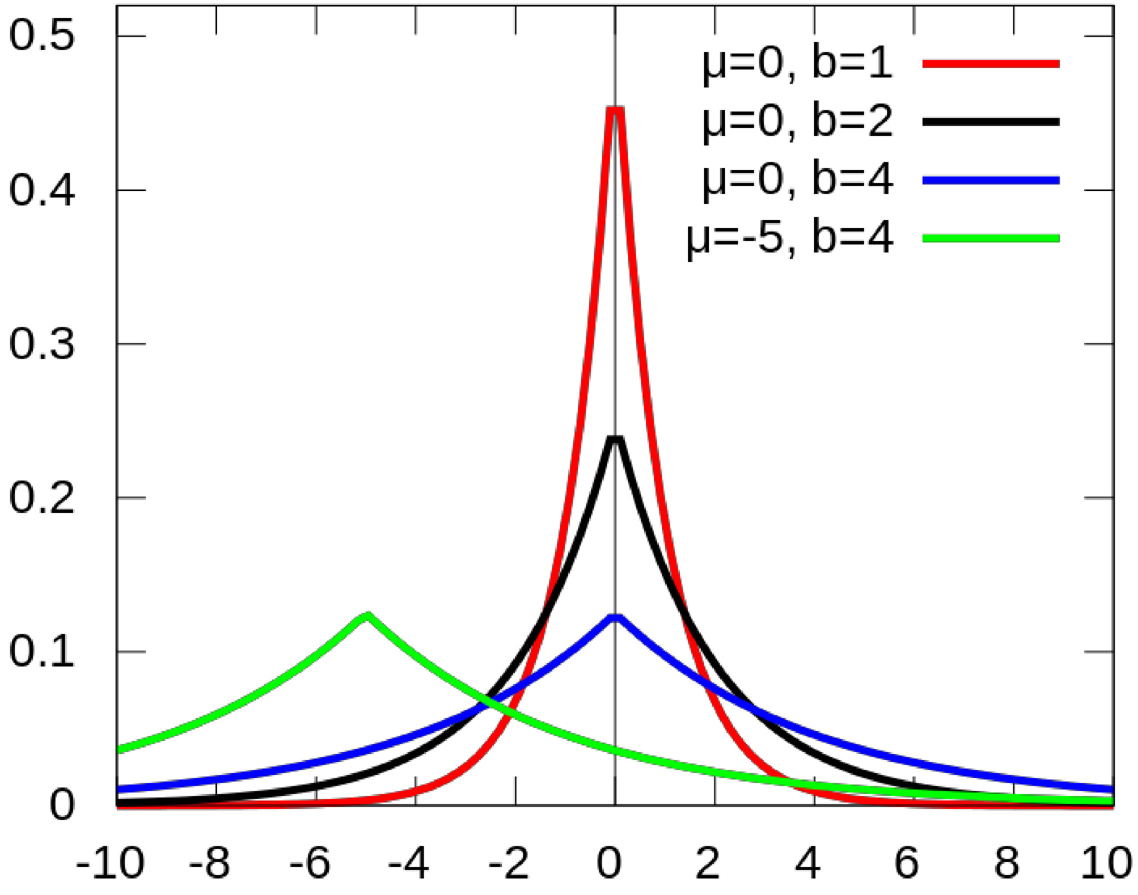
$$p(x) = \delta(x - \mu)$$



“Dirac density" of an idealized point mass or point charge -- a function that is equal to zero everywhere except for zero (integral over the entire real line is equal to one)

Laplace Distribution

$$\text{Laplace}(x; \mu, \gamma) = \frac{1}{2\gamma} \exp\left(-\frac{|x-\mu|}{\gamma}\right)$$



Bernoulli Distribution

The Bernoulli distribution is the “coin flip” distribution.

X is Bernoulli if its probability function is:

$$X = \begin{cases} 1 & w.p. \quad p \\ 0 & w.p. \quad 1-p \end{cases}$$

X=1 is usually interpreted as a “success.” E.g.:

- X=1 for heads in coin toss
- X=1 for male in survey
- X=1 for defective in a test of product
- X=1 for “made the sale” tracking performance

Bernoulli Distribution

$$P(\mathrm{x} = 1) = \phi$$

$$P(\mathrm{x} = 0) = 1 - \phi$$

$$P(\mathrm{x} = x) = \phi^x (1 - \phi)^{1-x}$$

$$\mathbb{E}_{\mathrm{x}}[\mathrm{x}] = \phi$$

$$\mathrm{Var}_{\mathrm{x}}(\mathrm{x}) = \phi(1 - \phi)$$

Can prove/derive each of these properties!

p is ϕ in these formulas!

Empirical Distribution

$$\hat{p}(\boldsymbol{x}) = \frac{1}{m}\sum_{i=1}^{m}\delta(\boldsymbol{x} - \boldsymbol{x}^{(i)})$$

An **empirical (Dirac) distribution function** is distribution function associated with empirical measure of a sample

Mixture Distributions

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i)$$

Gaussian mixture with
three components

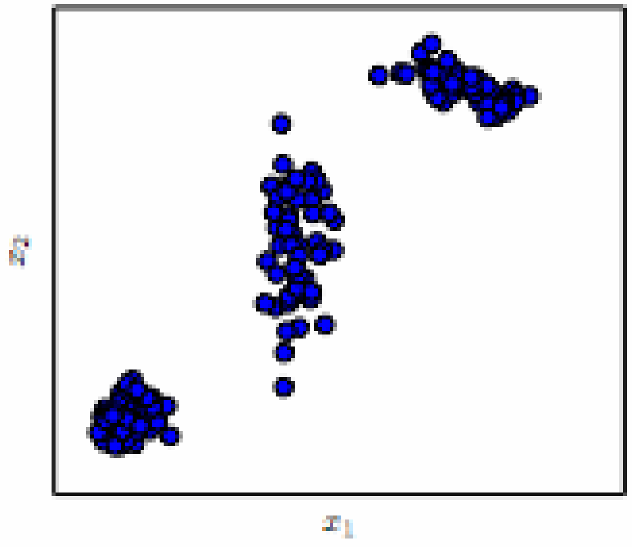


Figure 3.2

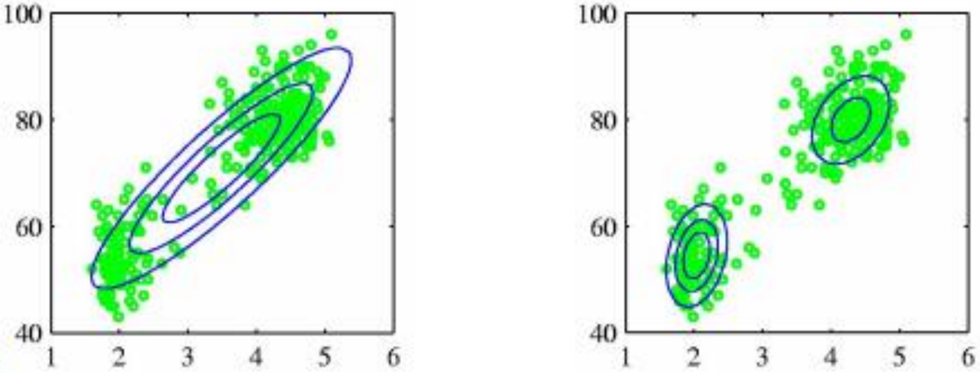
Mixtures of Gaussians

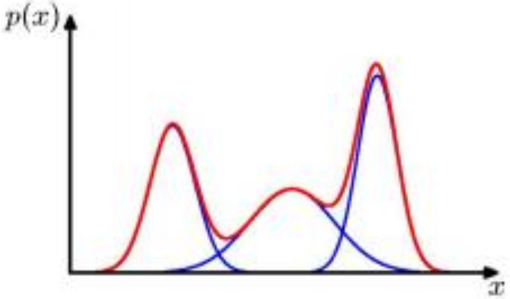
- Gaussian has limitations in modeling real data sets
- Old Faithful (Hydrothermal Geyser in Yellowstone)
  - 272 observations
  - Duration (mins, horiz axis) *vs* Time to next eruption (vertical axis)
  - Simple Gaussian unable to capture structure
  - Linear superposition of two Gaussians is better
- Linear combinations of Gaussians can give very complex densities

$$p(\mathrm{x}) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

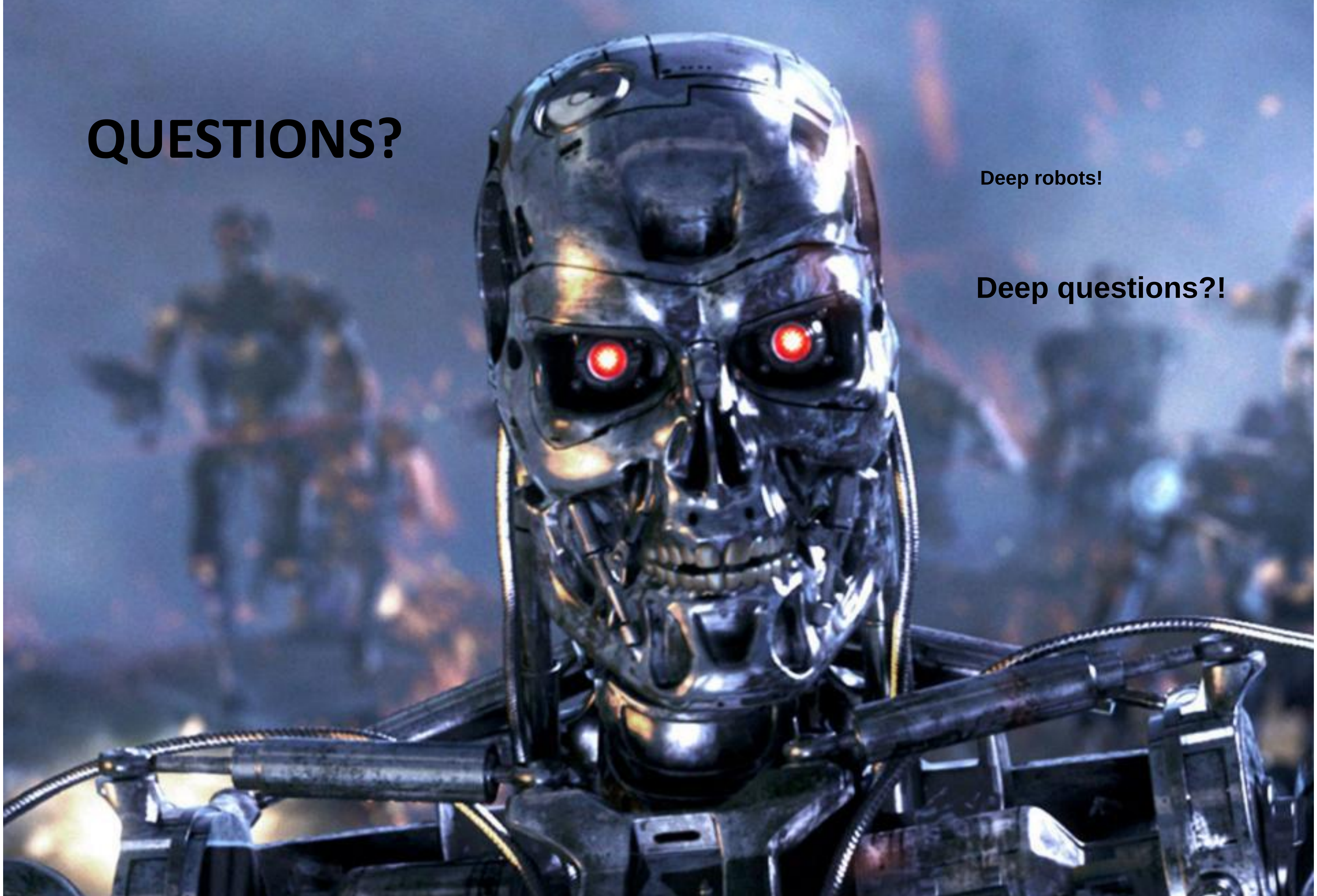
π_k are mixing coefficients that sum to one

- One –dimension
  - Three Gaussians in blue
  - Sum in red





QUESTIONS?



Deep robots!

Deep questions?!