Statistics and Fundamental Statistical Learning (II)

Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-335

2/9/2026

The Gaussian Distribution

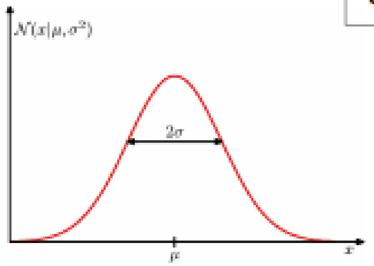
Carl Friedrich Gauss
1777-1855

- 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 $\mathbf{x}$, multivariate Gaussian

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

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

The Multivariate Gaussian Distribution

A p-dimensional random vector $\vec{X}$ has the *multivariate normal distribution* if it has the density function

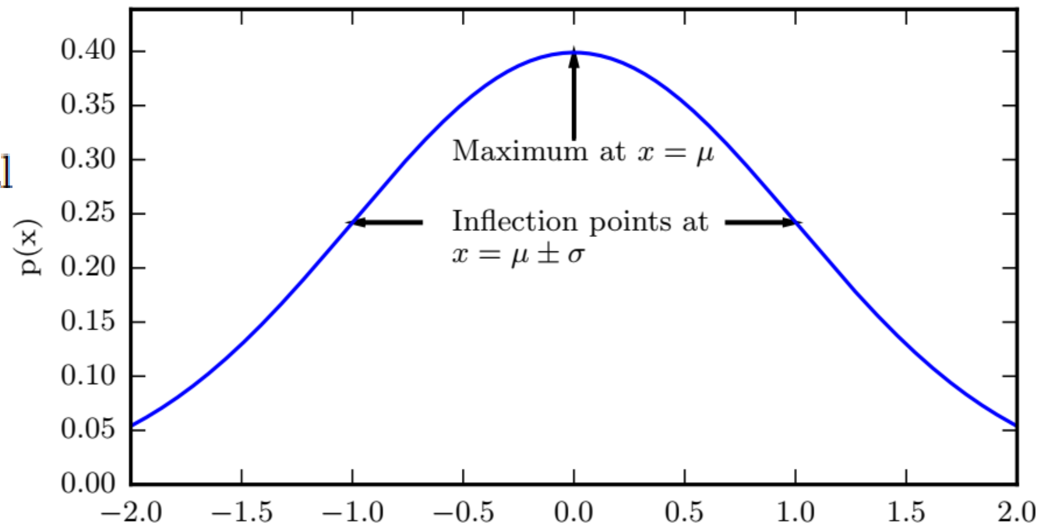
$$f(\vec{X}) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left(-\frac{1}{2}(\vec{X}-\vec{\mu})^T\Sigma^{-1}(\vec{X}-\vec{\mu})\right),$$

where $\vec{\mu}$ is a constant vector of dimension p and Σ is a $p \times p$ positive semi-definite which is invertible (called, in this case, *positive definite*). Then, $E\vec{X} = \vec{\mu}$ and $\mathrm{Cov}(\vec{X}) = \Sigma$.

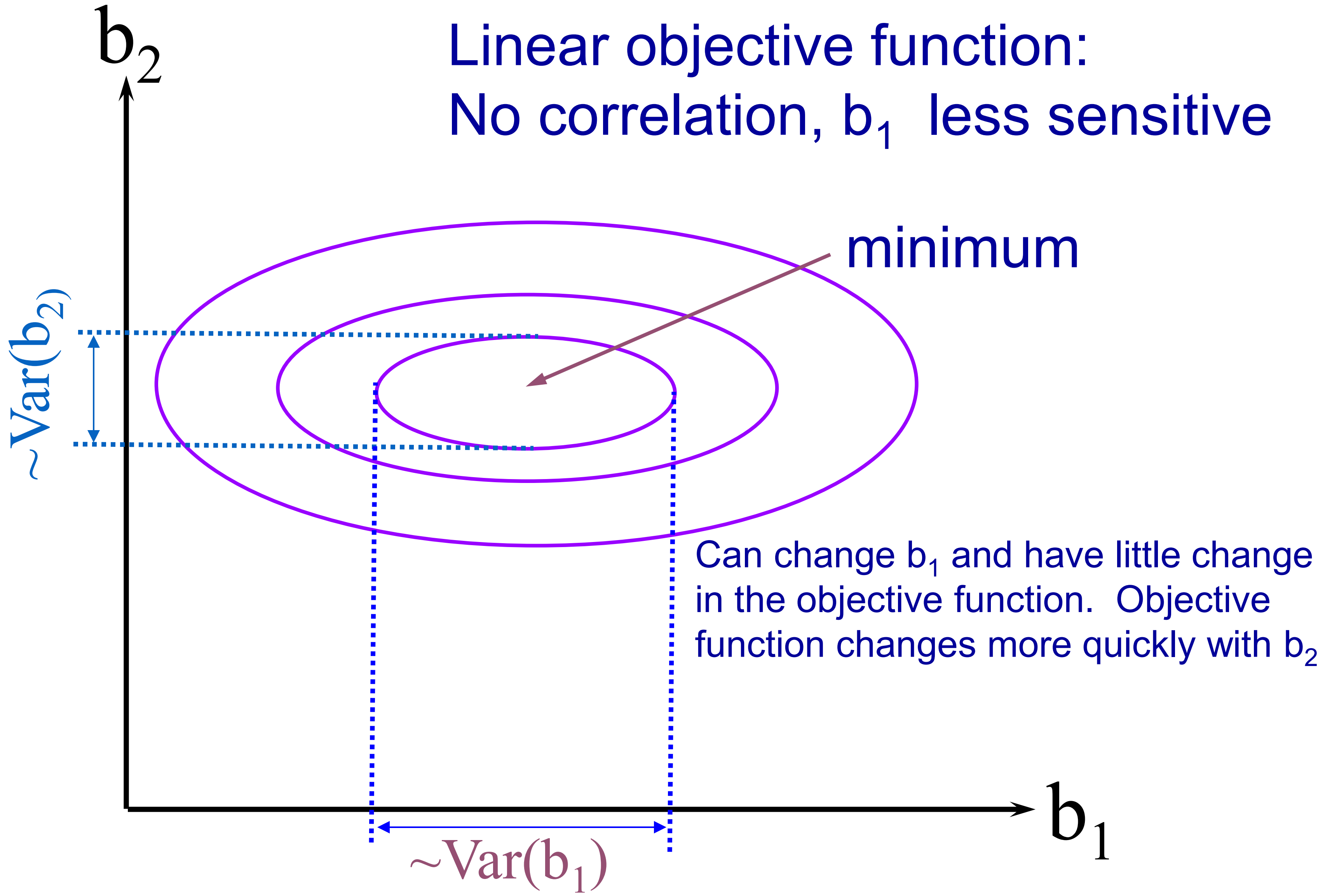
The *standard multivariate normal distribution* is obtained when $\vec{\mu} = 0$ and $\Sigma = I_p$, the $p \times p$ identity matrix:

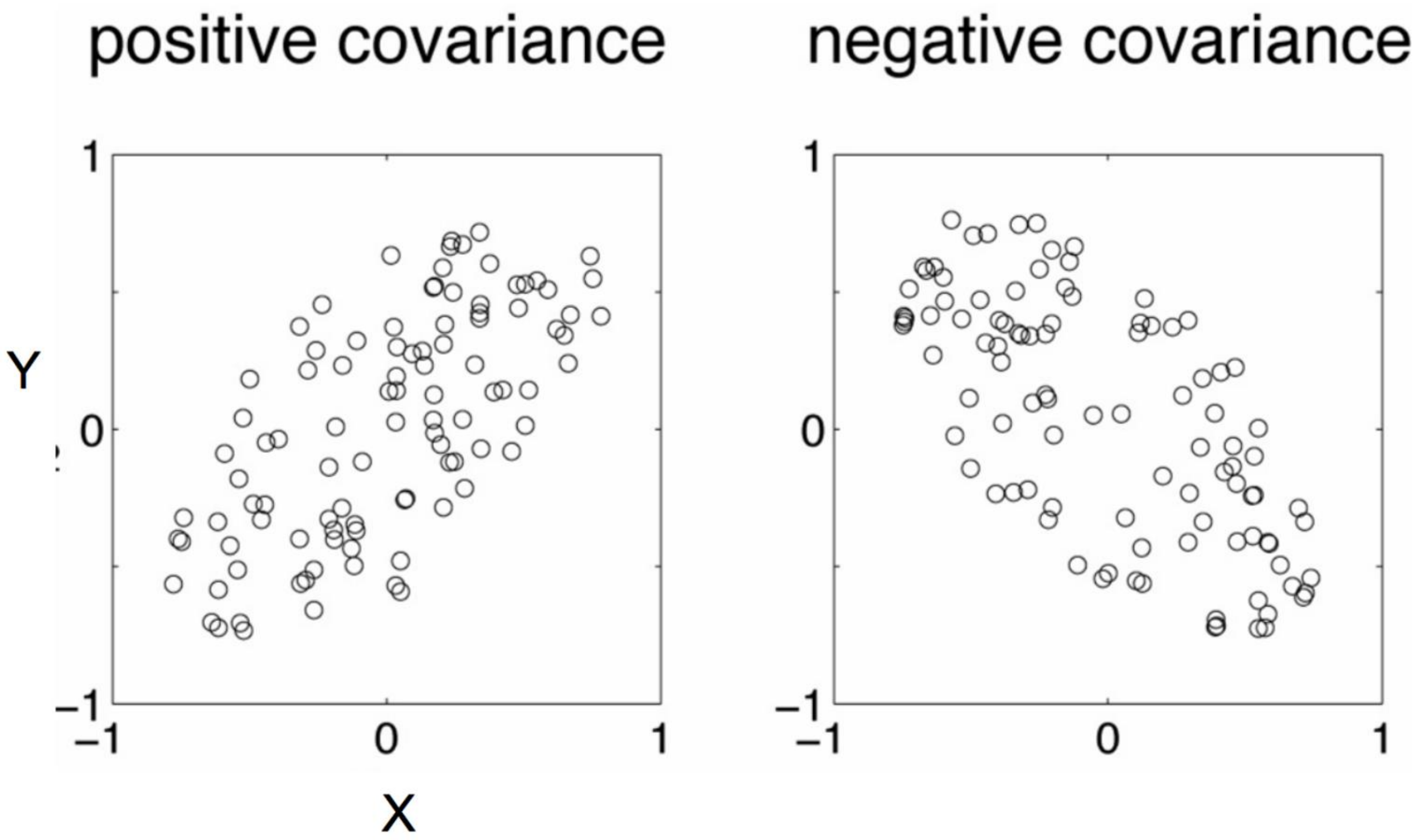
$$f(\vec{X}) = (2\pi)^{-p/2}\exp\left(-\frac{1}{2}\vec{X}^T\vec{X}\right).$$

This corresponds to the case where $X_1, \ldots, X_p$ are i.i.d. standard normal



Example: Parameter Variance & Covariance





Positive: Both dimensions increase or decrease together

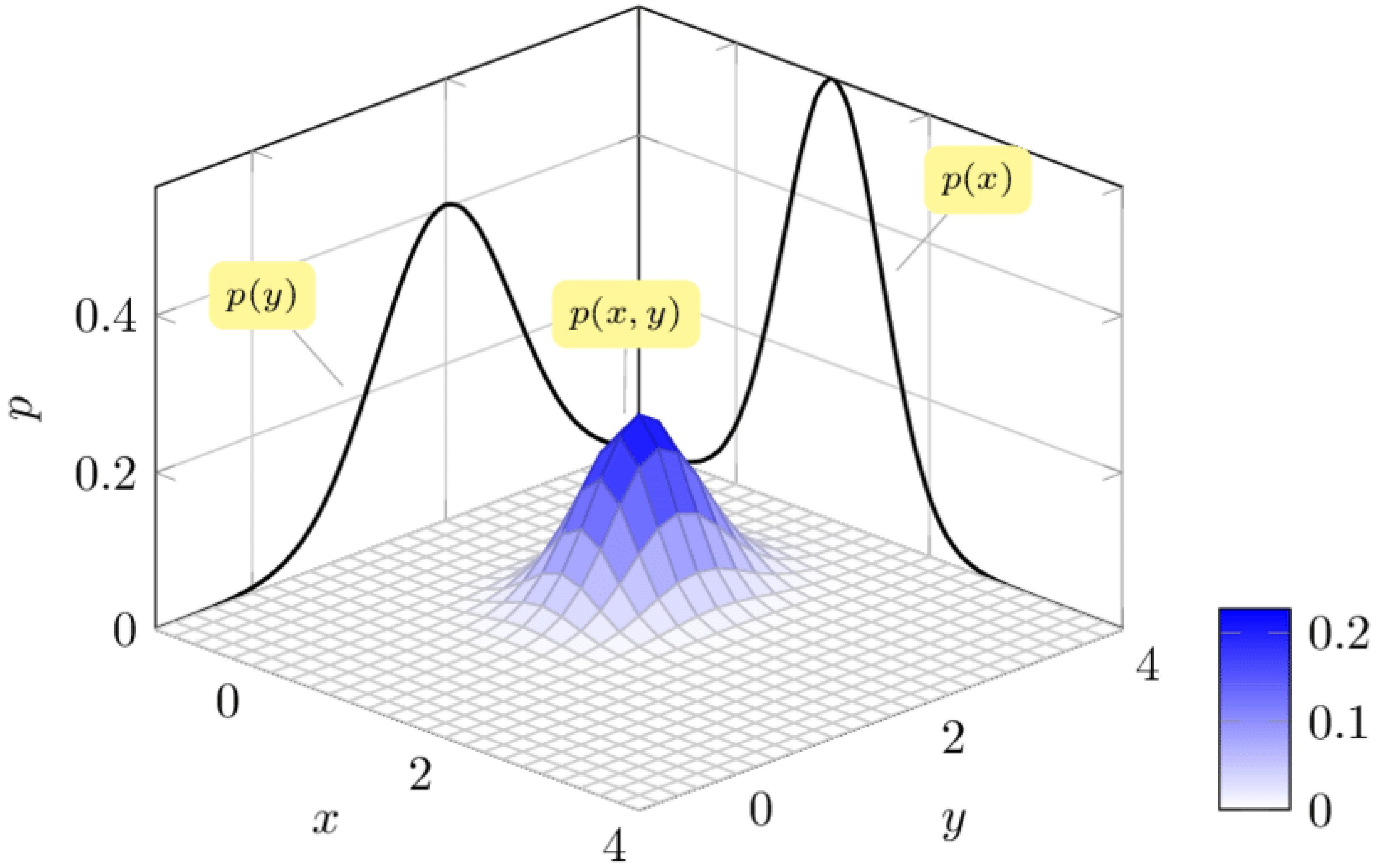
Negative: While one increases the other decreases

Covariance

- Used to find relationships between dimensions in high dimensional data sets

$$q_{jk} = \frac{1}{N} \sum_{i=1}^{N} (X_{ij} - E(X_j))(X_{ik} - E(X_k))$$

The sample mean



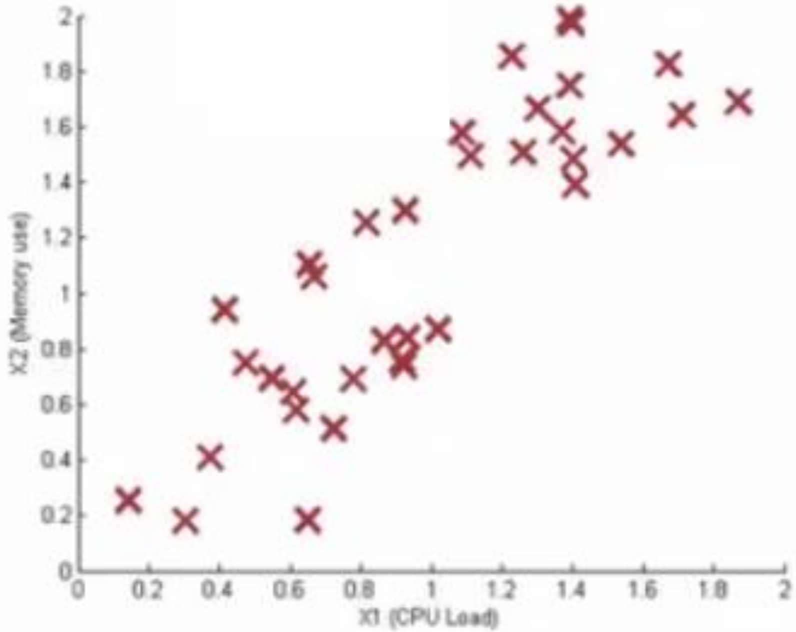
$p(y)$
$p(x,y)$
$p(x)$
p
0.4
0.2
0
0
2
4
x
0
2
4
y
0.2
0.1
0

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

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

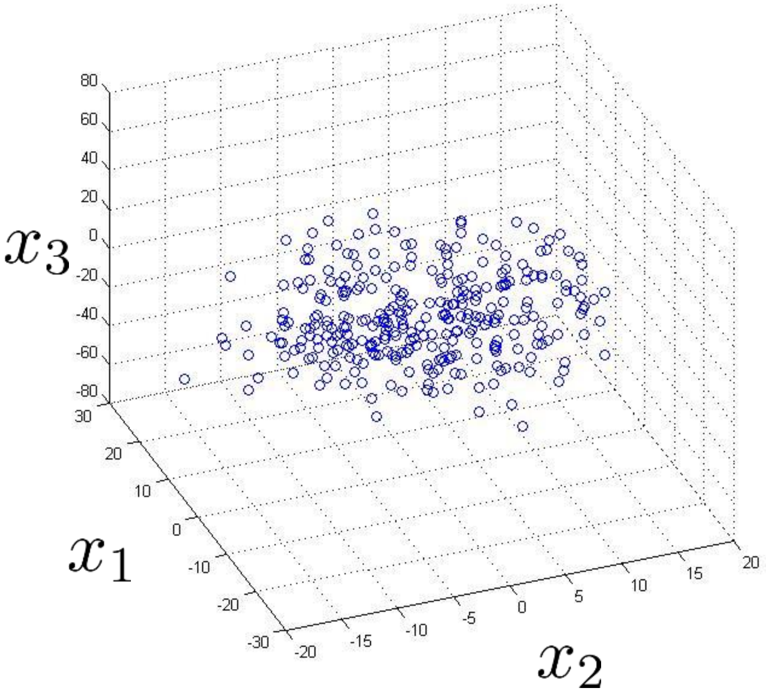
Flag an anomaly if $p(x) < \varepsilon$

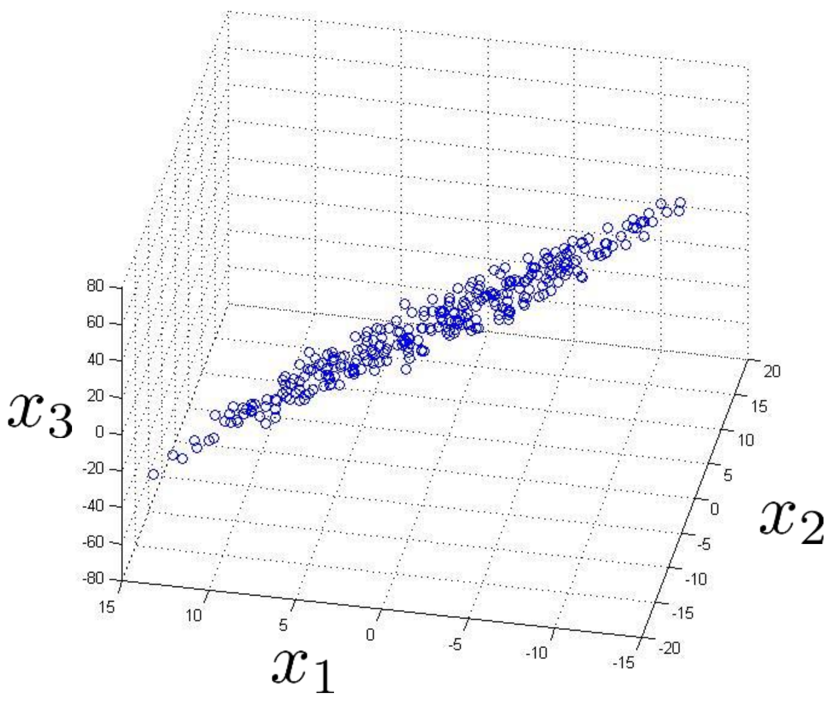
Why? Well, Dimensionality Reduction…

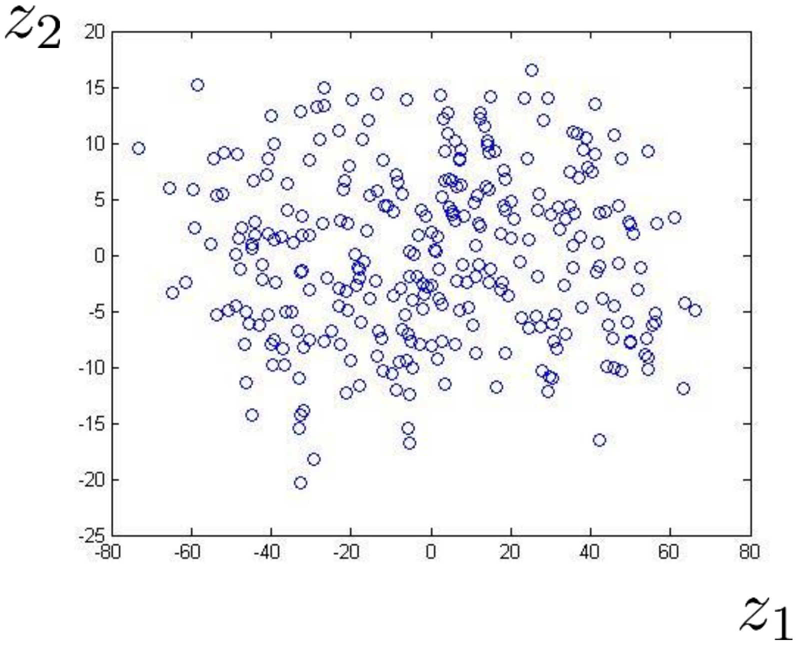
- PCA (Principal Component Analysis):
  - Find projection that maximize the variance (based on Gaussian assumptions)
- ICA (Independent Component Analysis):
  - Similar to PCA except assumes non-Gaussian features
- Multidimensional Scaling:
  - Find projection that best preserves inter-point distances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

Data Compression

Reduce data from 3D to 2D





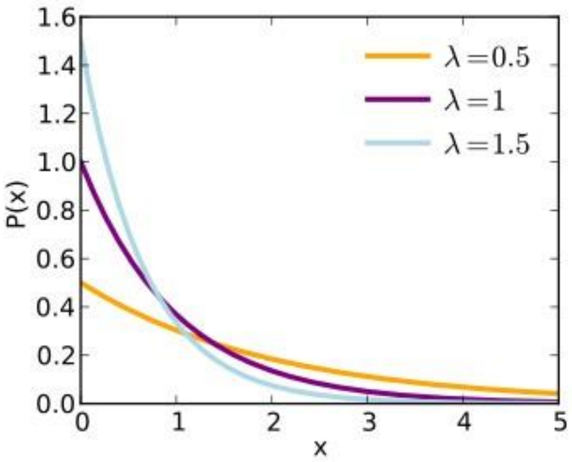


Andrew Ng

More Distributions

Exponential:

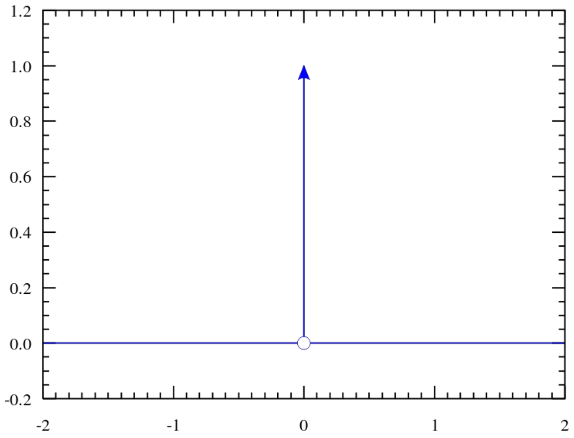
$$p(x;\lambda) = \lambda \mathbf{1}_{x\geq 0} \exp\left(-\lambda x\right).$$



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

Dirac

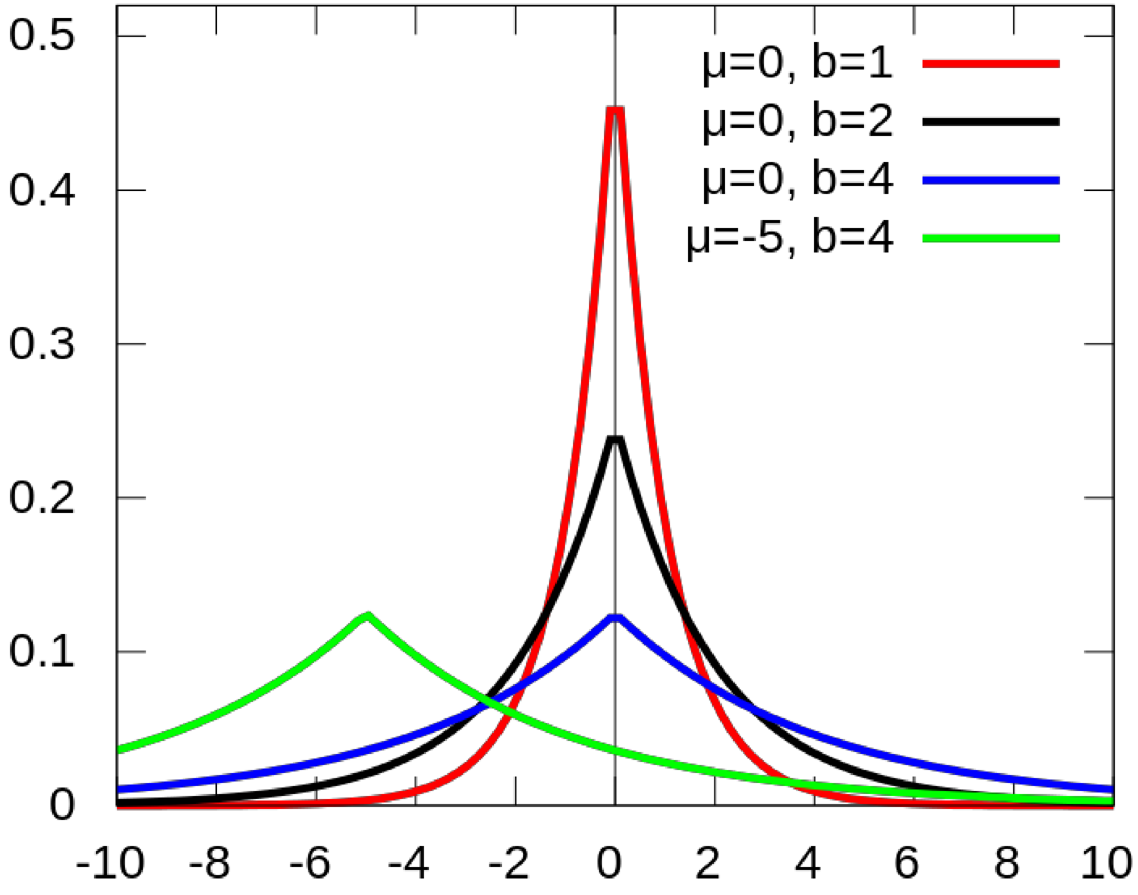
$$p(x) = \delta(x-\mu)$$



“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

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0 \end{cases}$$

such that:

$$\int_{-\infty}^{\infty} \delta(x)dx = 1$$

Dirac delta function

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

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

Mixture Distributions

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

Gaussian mixture with
three components

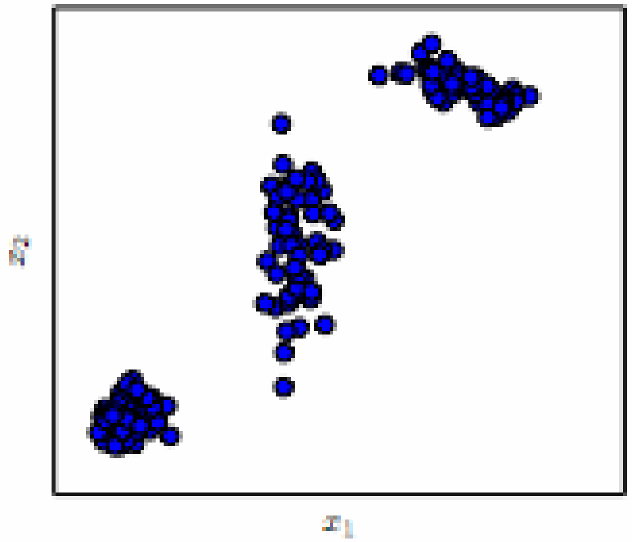
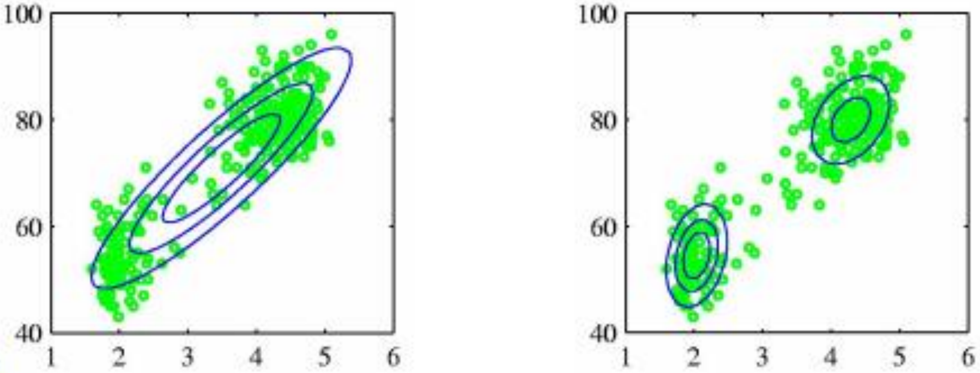


Figure 3.2

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

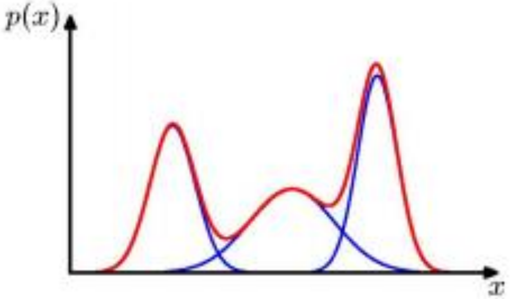


- Linear combinations of Gaussians can give very complex densities

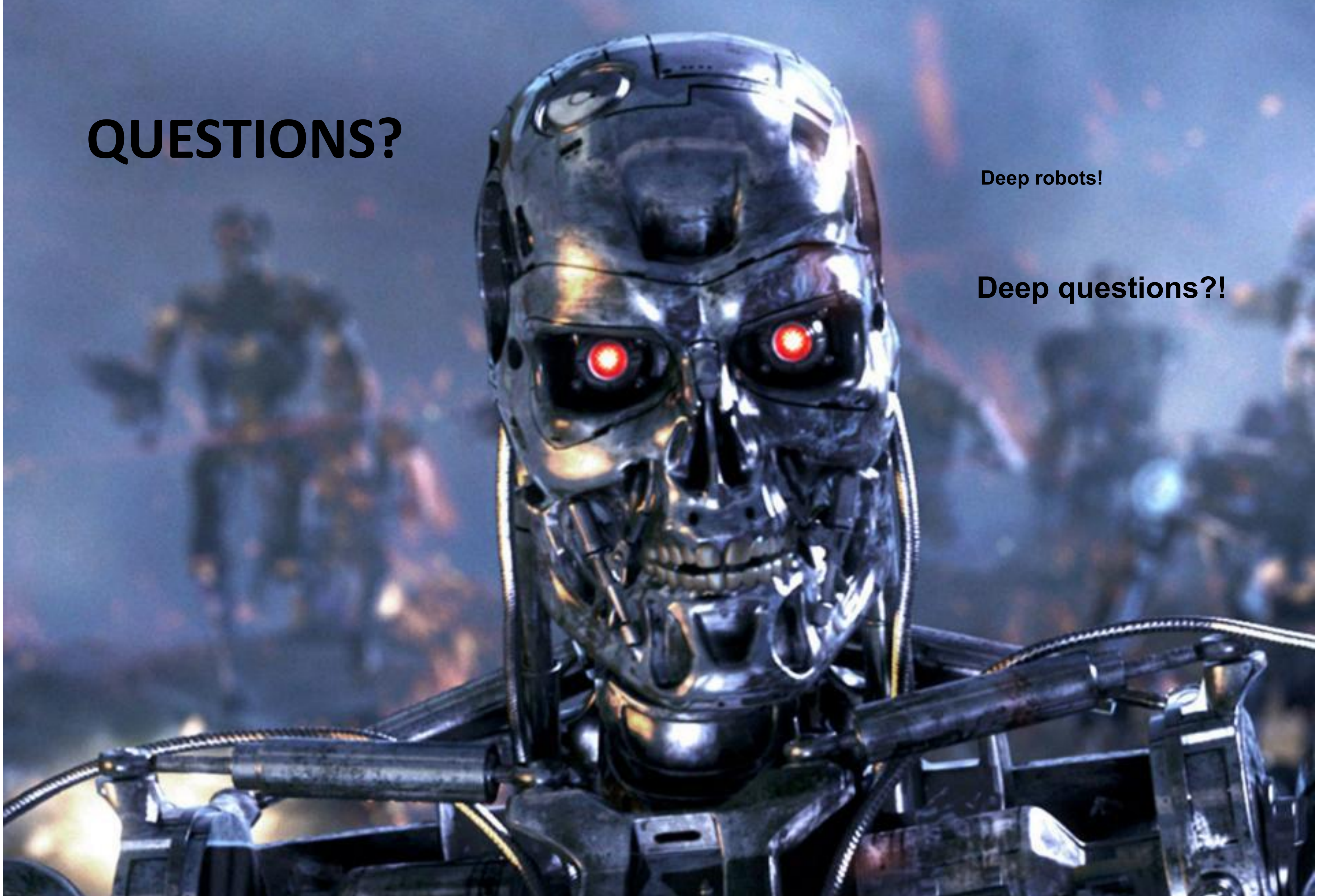
$$p(\mathrm{x}) = \sum_{k=1}^{K} \pi_k N(x \mid \mu_k, \Sigma_k)$$

π_k are mixing coefficients that sum to one

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



We will build this model later in this class!


QUESTIONS?
Deep robots!
Deep questions?!