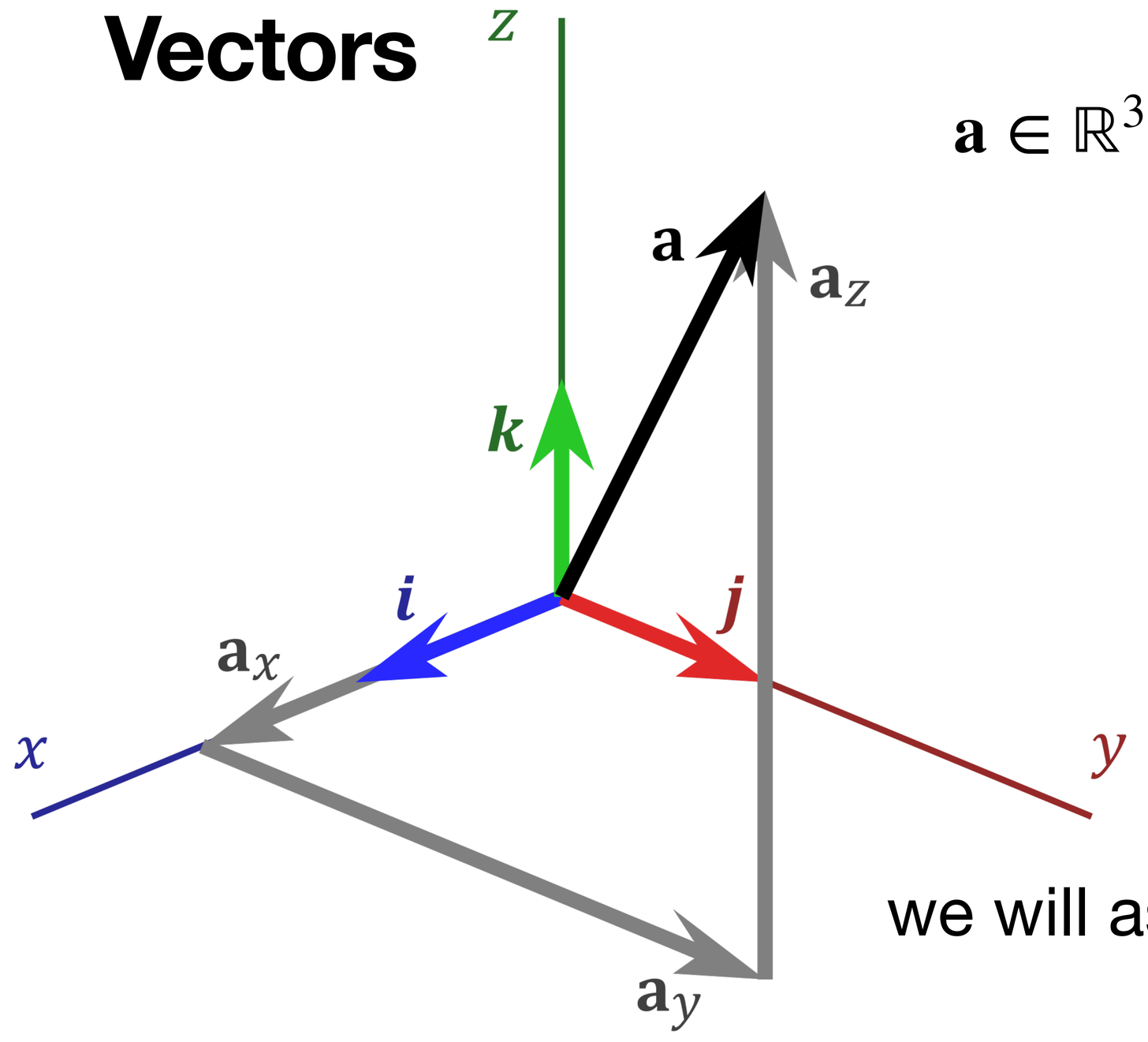
Linear algebra and multivariable Gaussians

Pre-lecture 4 video

Vector operations we need

- Vector addition
- Vector multiplication
  - Vector with scalar
  - Between two vectors to produce a scalar (dot product)
  - ~~Between two vectors to produce a vector (cross product)~~

Vectors



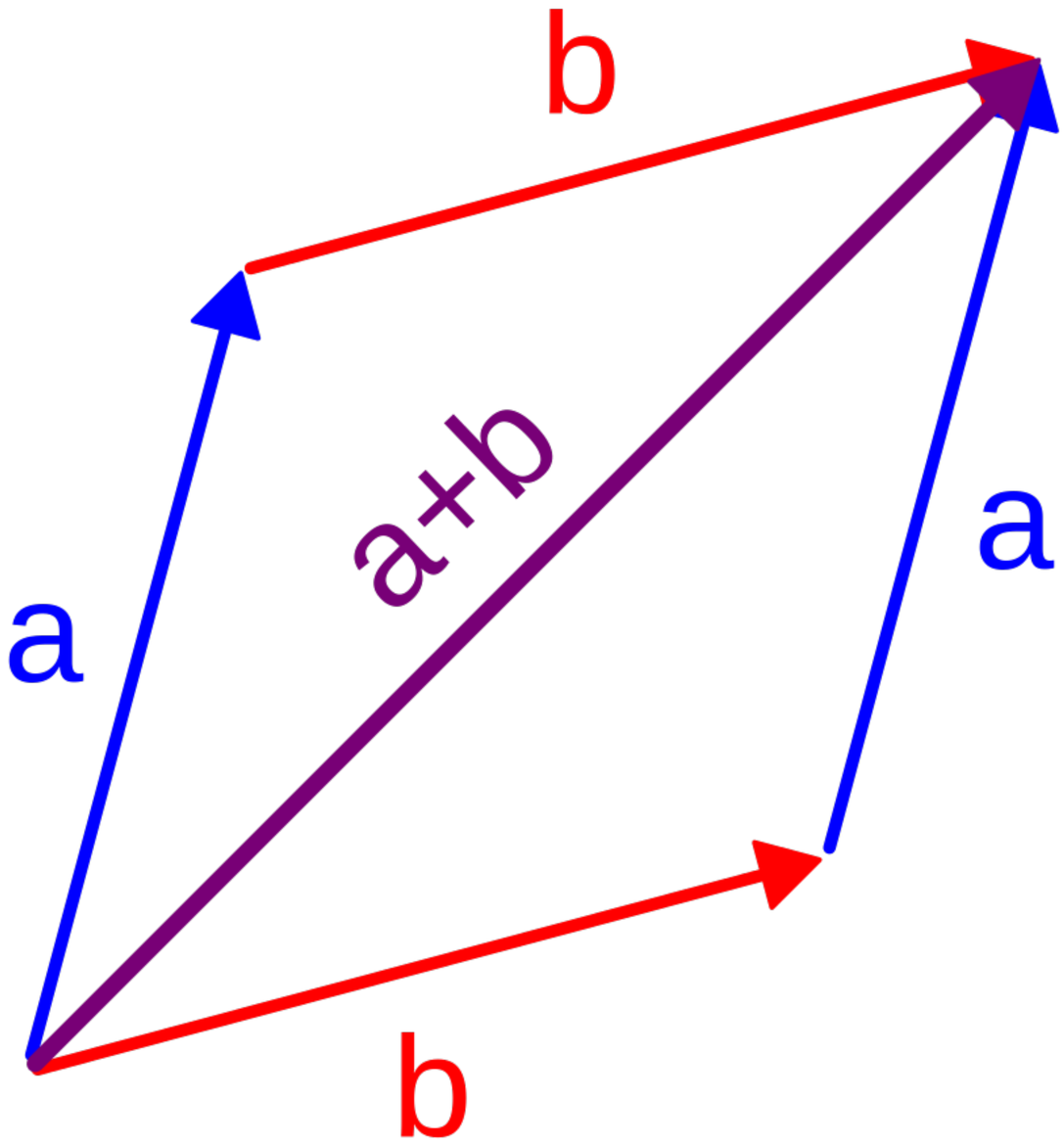
$$\mathbf{a} \in \mathbb{R}^3 \qquad \mathbf{a} = \begin{bmatrix} a_x \\ a_y \\ a_z \end{bmatrix}$$

$$\|\mathbf{a}\|_2 = \sqrt{a_x^2 + a_y^2 + a_z^2}$$

$$\|\mathbf{a}\|_n = \left(a_x^n + a_y^n + a_z^n \right)^{1/n}$$

we will assume that $\|\mathbf{a}\|$ means $\|\mathbf{a}\|_2$

Vector addition

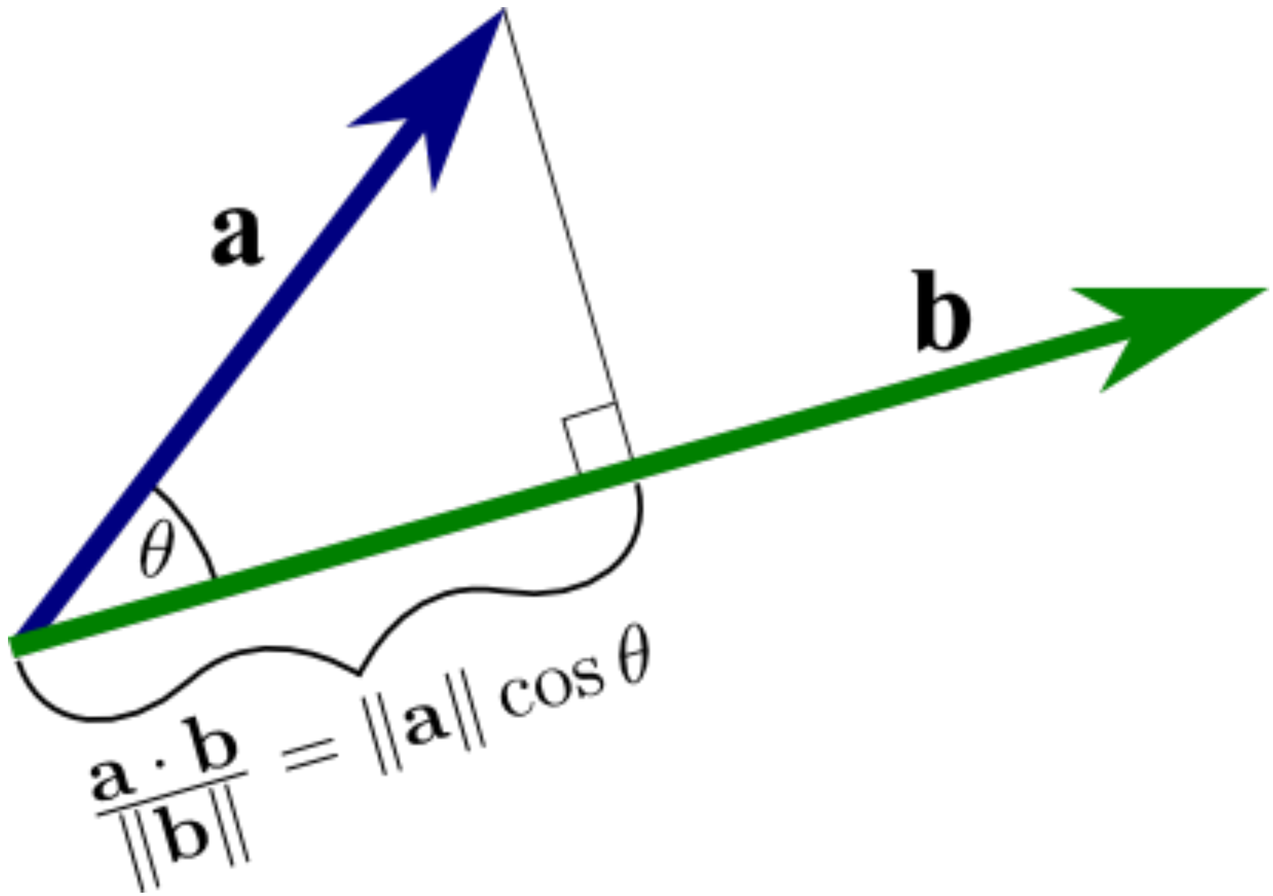


$$\mathbf{a}, \mathbf{b} \in \mathbb{R}^2$$

$$\mathbf{a} = \begin{bmatrix} a_x \\ a_y \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_x \\ b_y \end{bmatrix}$$

$$\mathbf{a} + \mathbf{b} = \begin{bmatrix} a_x + b_x \\ a_y + b_y \end{bmatrix}$$

Dot product - a scalar projection



$$\mathbf{a}\cdot\mathbf{b} = \|\mathbf{a}\|\,\|\mathbf{b}\|\,\cos\theta$$

$$\mathbf{a}\cdot\mathbf{b} \equiv \langle\mathbf{a},\mathbf{b}\rangle$$

$$\mathbf{a}\cdot\mathbf{b} \equiv \mathbf{a}^{\mathrm{T}}\mathbf{b}$$

Matrix multiplication

$$\mathbf{a} = (a_1, a_2, \ldots, a_n) \qquad \mathbf{a}^\mathsf{T} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

$$\mathbf{ab} = a_1 b_1 + a_2 b_2 \ldots a_n b_n \qquad \mathbf{ba} = \begin{pmatrix} b_1 a_1 & b_1 a_2 & \ldots & b_1 a_n \\ b_2 a_1 & b_2 a_2 & \ldots & b_2 a_n \\ \vdots & & \ddots & \\ b_3 a_1 & b_3 a_2 & \ldots & b_3 a_n \end{pmatrix}$$

Matrix multiplication

Matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \qquad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}$$

Determinant

69 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

This article is about mathematics. For determinants in epidemiology, see Risk factor. For determinants in immunology, see Epitope.

In mathematics, the **determinant** is a scalar value that is a function of the entries of a square matrix. The determinant of a matrix A is commonly denoted $\det(A)$, $\det A$, or $|A|$. Its value characterizes some properties of the matrix and the linear map represented by the matrix. In particular, the determinant is nonzero if and only if the matrix is invertible and the linear map represented by the matrix is an isomorphism. The determinant of a product of matrices is the product of their determinants.

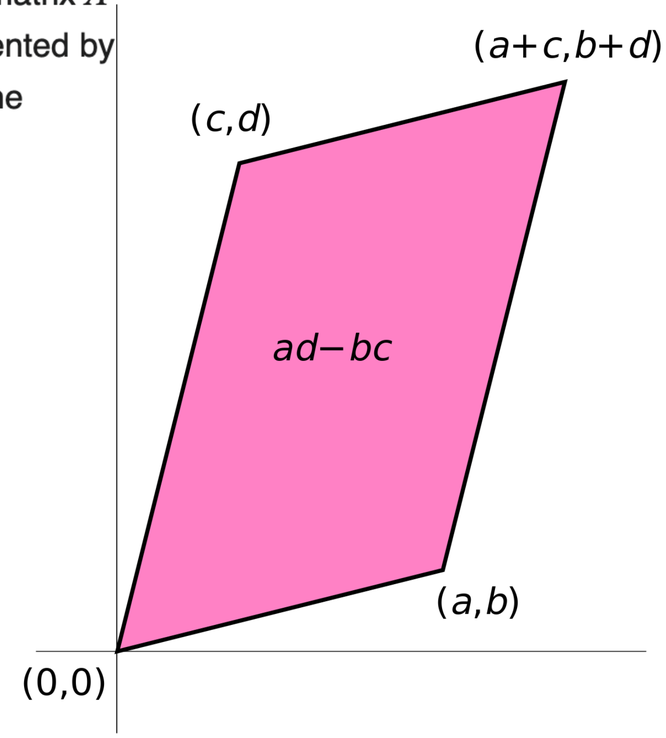
The determinant of a 2×2 matrix is

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc,$$

and the determinant of a 3×3 matrix is

$$\begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = aei + bfg + cdh - ceg - bdi - afh.$$

Geometric interpretation



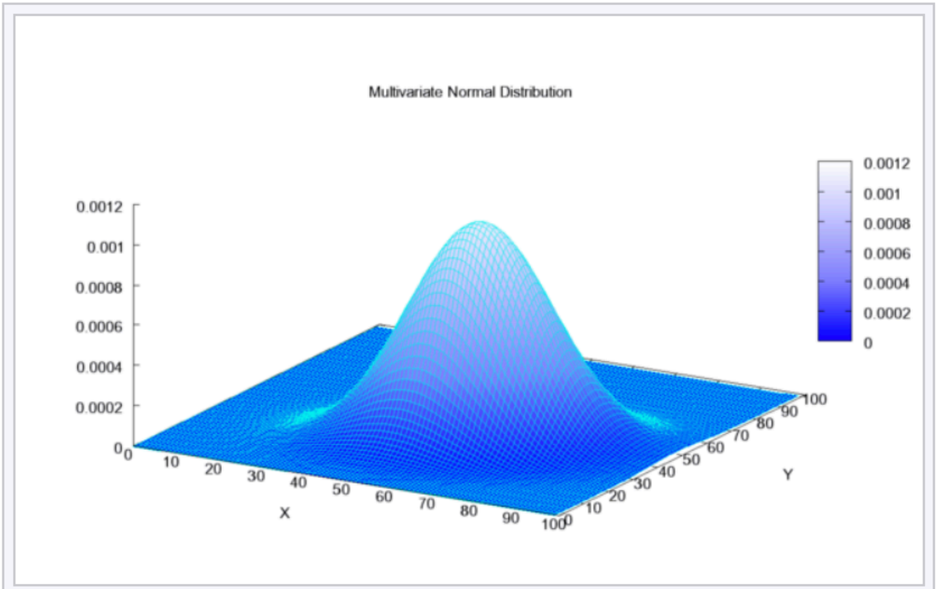
Multivariate Normal distribution

Density function [edit]

Non-degenerate case [edit]

The multivariate normal distribution is said to be "non-degenerate" when the symmetric covariance matrix $\boldsymbol{\Sigma}$ is positive definite. In this case the distribution has density[5]

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$



Bivariate normal joint density

where $\mathbf{x}$ is a real *k*-dimensional column vector and $|\boldsymbol{\Sigma}| \equiv \det \boldsymbol{\Sigma}$ is the determinant of $\boldsymbol{\Sigma}$, also known as the generalized variance. The equation above reduces to that of the univariate normal distribution if $\boldsymbol{\Sigma}$ is a 1×1 matrix (i.e. a single real number).

Vectors represent variables

[Position x, Position y, Position z, Velocity x, Velocity y, Velocity z]

[House price, Year built, Square footage, # Bedrooms, # Bathrooms]

[Make, Model, Year built, Engine displacement, Miles per gallon, Color]

[Weight, # legs, lays eggs?, flys?, ….]

	fly?	laying eggs?	weight (lb)
sparrow	yes	yes	0.087
chipmunk	no	no	0.19
bat	yes	no	0.09

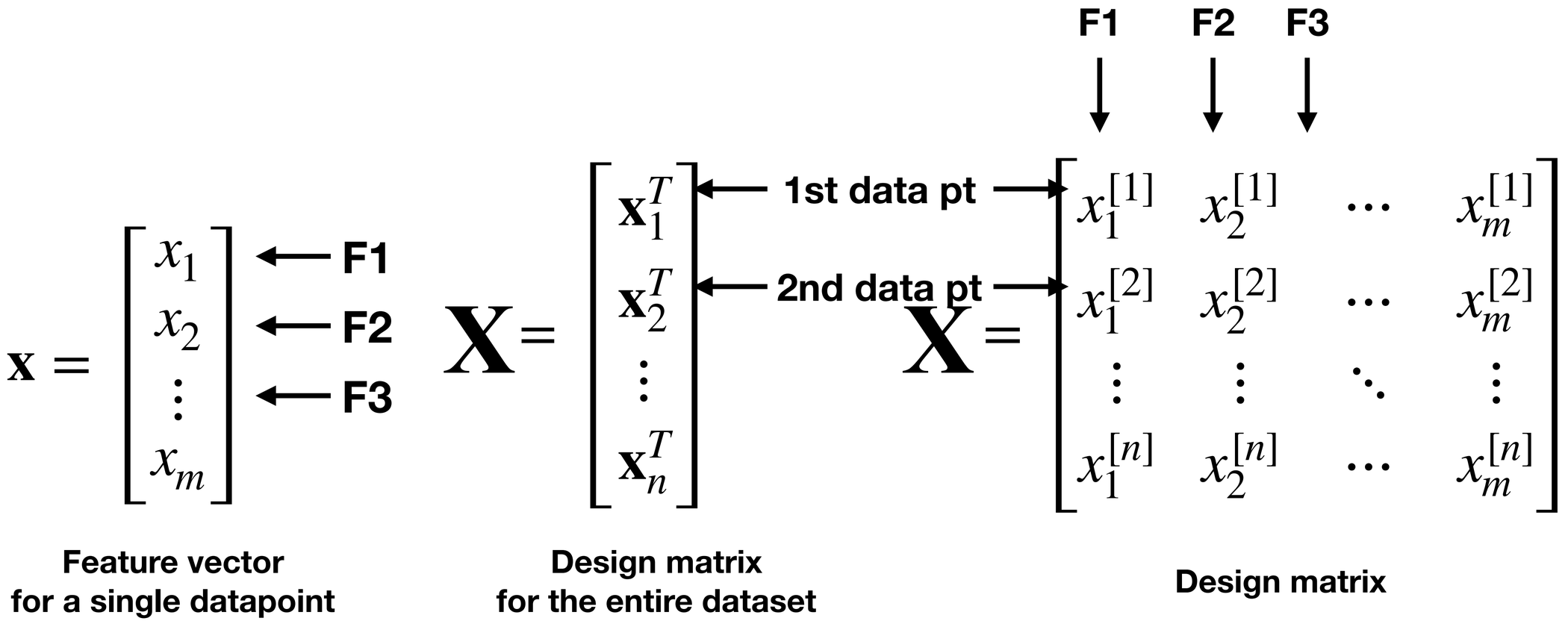
Feature representation (category encoded)

$$\text{sparrow} = \begin{pmatrix} \text{True} \\ \text{True} \\ 0.087 \end{pmatrix} \qquad \text{chipmunk} = \begin{pmatrix} \text{False} \\ \text{False} \\ 0.19 \end{pmatrix} \qquad \text{bat} = \begin{pmatrix} \text{True} \\ \text{False} \\ 0.09 \end{pmatrix}$$

Feature representation (one-hot encoded)

$$\text{sparrow} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0.087 \end{pmatrix} \qquad \text{chipmunk} = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0.19 \end{pmatrix} \qquad \text{bat} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0.09 \end{pmatrix}$$

Data Representation



Estimation II

Lecture 5

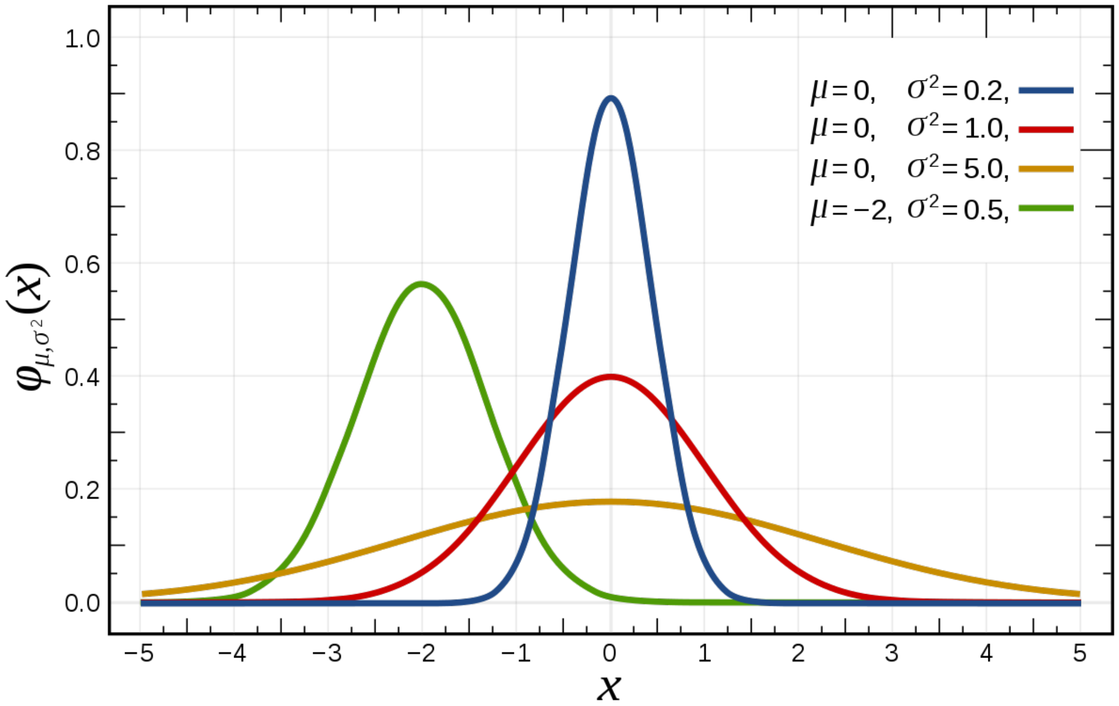
Announcements

- Common assignment problems
  - BEFORE you make conda env
    - Install LaTeX (MikTeX for Windows or MacTeX for MacOS)
    - Install Pandoc
    - THEN make the conda env
  - Important aspects for good Markdown LaTeX
    - List after math env will not work … why??? Dunno but it doesn’t
    - $$ $$ not $ $
    - \aligned not \align
    - File -> export -> PDF via LaTeX to debug your own markdown

Normal distribution

- “Normal” because its very common
- Central limit theorem: sum of a large number of independent RVs is normally distributed
- Analytically tractable
- Exponential family
- Maximum entropy of all distributions with a given mean and variance

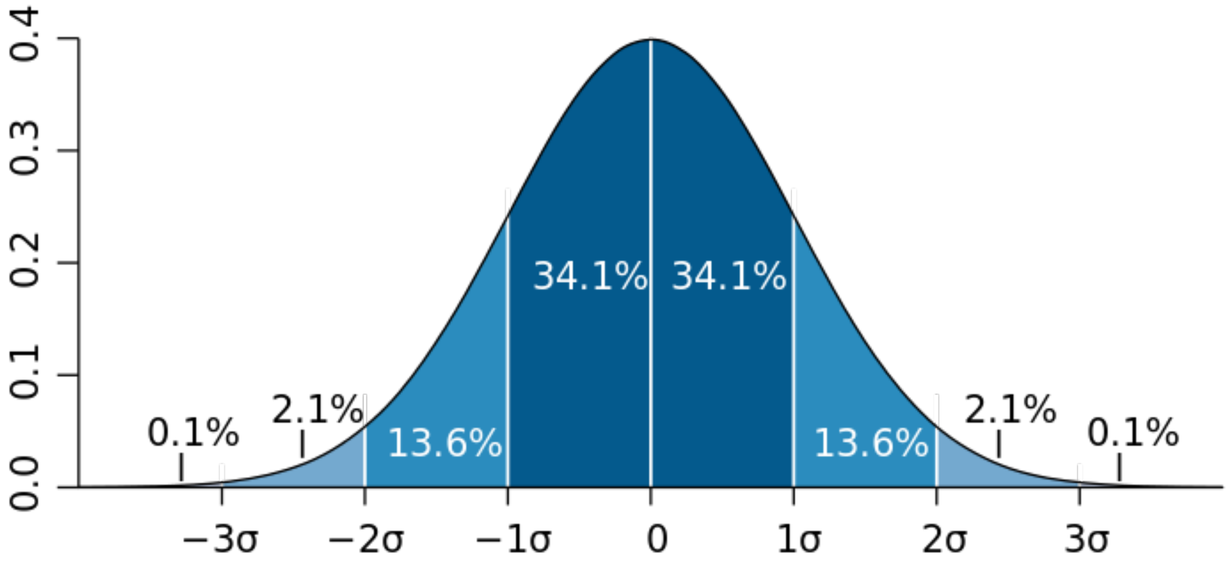
$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Normal distribution

- Exponential decay with distance squared from μ
- Distance is normalized in units of σ
- Constant in front normalizes PDF to 1

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Bayes with Normals

Suppose we have a measurement $x \sim N(\theta, \sigma^2)$ where the variance σ^2 is known. That is, the mean θ is our unknown parameter of interest and we are given that the likelihood comes from a normal distribution with variance σ^2. If we choose a normal prior pdf

$$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$$

then the posterior pdf is also normal: $f(\theta|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ where

$$\frac{\mu_{\text{post}}}{\sigma^2_{\text{post}}} = \frac{\mu_{\text{prior}}}{\sigma^2_{\text{prior}}} + \frac{x}{\sigma^2}, \qquad \frac{1}{\sigma^2_{\text{post}}} = \frac{1}{\sigma^2_{\text{prior}}} + \frac{1}{\sigma^2} \tag{1}$$

The following form of these formulas is easier to read and shows that μ_{post} is a weighted average between μ_{prior} and the data x.

$$a = \frac{1}{\sigma^2_{\text{prior}}} \qquad b = \frac{1}{\sigma^2}, \qquad \mu_{\text{post}} = \frac{a\mu_{\text{prior}} + bx}{a+b}, \qquad \sigma^2_{\text{post}} = \frac{1}{a+b}. \tag{2}$$

hypothesis	data	prior	likelihood	posterior
θ	x	$f(\theta) \sim \mathrm{N}(\mu_{\text{prior}}, \sigma^2_{\text{prior}})$ $= c_1 \exp\left(\frac{-(\theta-\mu_{\text{prior}})^2}{2\sigma^2_{\text{prior}}}\right)$	$\phi(x\|\theta) \sim \mathrm{N}(\theta, \sigma^2)$ $= c_2 \exp\left(\frac{-(x-\theta)^2}{2\sigma^2}\right)$	$f(\theta\|x) \sim \mathrm{N}(\mu_{\text{post}}, \sigma^2_{\text{post}})$ $= c_3 \exp\left(\frac{-(\theta-\mu_{\text{post}})^2}{2\sigma^2_{\text{post}}}\right)$

Bayes with Normals

Example 2. Suppose we have prior $\theta \sim \mathrm{N}(4, 8)$, and likelihood function likelihood $x \sim \mathrm{N}(\theta, 5)$. Suppose also that we have one measurement $x_1 = 3$. Show the posterior distribution is normal.

answer: We will show this by grinding through the algebra which involves completing the square.

$$\text{prior: } f(\theta) = c_1\, \mathrm{e}^{-(\theta-4)^2/16}; \qquad \text{likelihood: } \phi(x_1|\theta) = c_2\, \mathrm{e}^{-(x_1-\theta)^2/10} = c_2\, \mathrm{e}^{-(3-\theta)^2/10}$$

We multiply the prior and likelihood to get the posterior:

$$\begin{aligned} f(\theta|x_1) &= c_3\, \mathrm{e}^{-(\theta-4)^2/16}\, \mathrm{e}^{-(3-\theta)^2/10} \\ &= c_3 \exp\left(-\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10}\right) \end{aligned}$$

We complete the square in the exponent

$$\begin{aligned} -\frac{(\theta-4)^2}{16} - \frac{(3-\theta)^2}{10} &= -\frac{5(\theta-4)^2 + 8(3-\theta)^2}{80} \\ &= -\frac{13\theta^2 - 88\theta + 152}{80} \\ &= -\frac{\theta^2 - \frac{88}{13}\theta + \frac{152}{13}}{80/13} \\ &= -\frac{(\theta - 44/13)^2 + 152/13 - (44/13)^2}{80/13}. \end{aligned}$$

Therefore the posterior is

$$f(\theta|x_1) = c_3\, \mathrm{e}^{-\frac{(\theta-44/13)^2 + 152/13 - (44/13)^2}{80/13}} = c_4\, \mathrm{e}^{-\frac{(\theta-44/13)^2}{80/13}}.$$

This has the form of the pdf for $\mathrm{N}(44/13, 40/13)$. QED

For practice we check this against the formulas (2).

$$\mu_{\text{prior}} = 4, \quad \sigma^2_{\text{prior}} = 8, \quad \sigma^2 = 5 \;\Rightarrow\; a = \frac{1}{8}, \quad b = \frac{1}{5}.$$

Therefore

$$\begin{aligned} \mu_{\text{post}} &= \frac{a\mu_{\text{prior}} + bx}{a+b} = \frac{44}{13} = 3.38 \\ \sigma^2_{\text{post}} &= \frac{1}{a+b} = \frac{40}{13} = 3.08. \end{aligned}$$

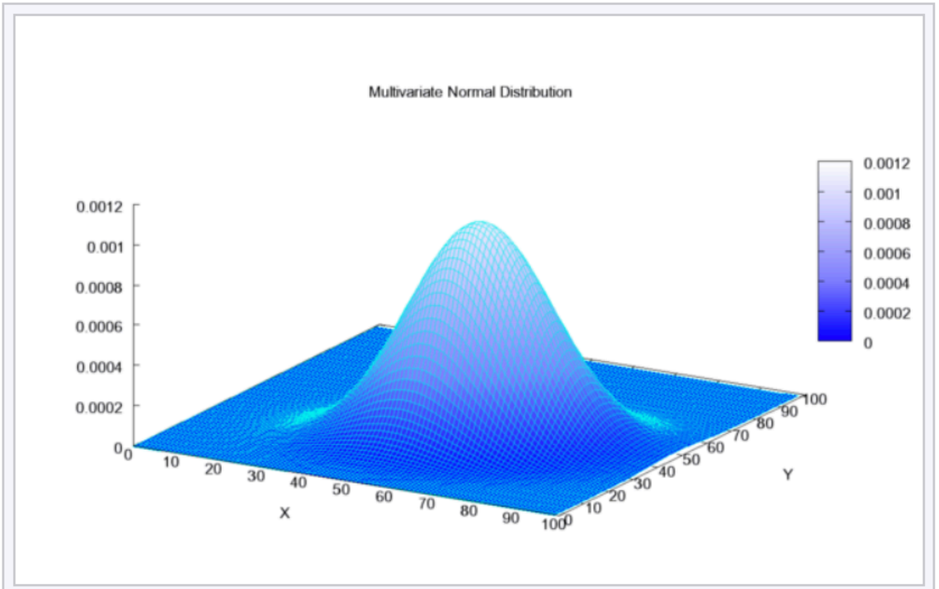
Multivariate Normal distribution

Density function [edit]

Non-degenerate case [edit]

The multivariate normal distribution is said to be "non-degenerate" when the symmetric covariance matrix $\boldsymbol{\Sigma}$ is positive definite. In this case the distribution has density[5]

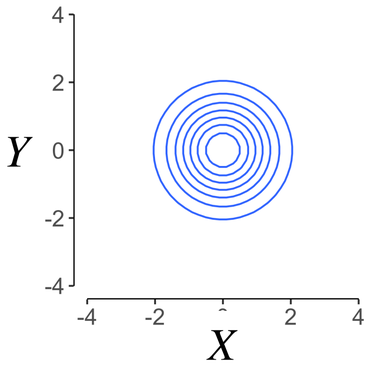
$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$



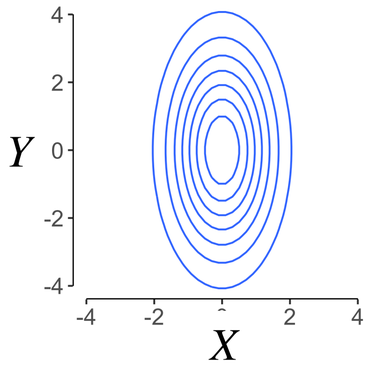
Bivariate normal joint density

where $\mathbf{x}$ is a real *k*-dimensional column vector and $|\boldsymbol{\Sigma}| \equiv \det \boldsymbol{\Sigma}$ is the determinant of $\boldsymbol{\Sigma}$, also known as the generalized variance. The equation above reduces to that of the univariate normal distribution if $\boldsymbol{\Sigma}$ is a 1×1 matrix (i.e. a single real number).

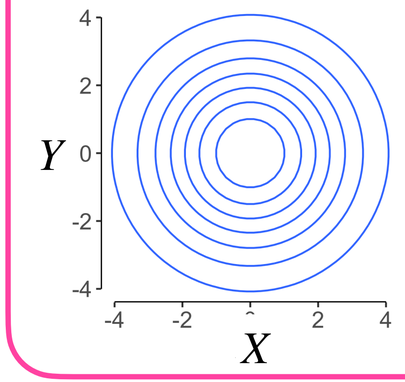
$\rho = 0, \sigma_X = \sigma_Y = 1$



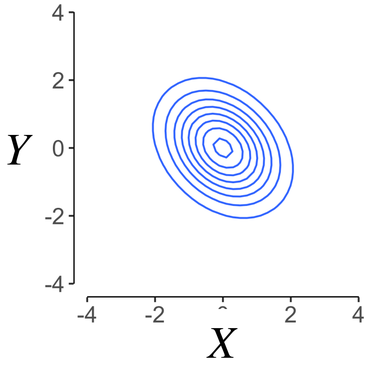
$\rho = 0, \sigma_X = 1, \sigma_Y = 2$



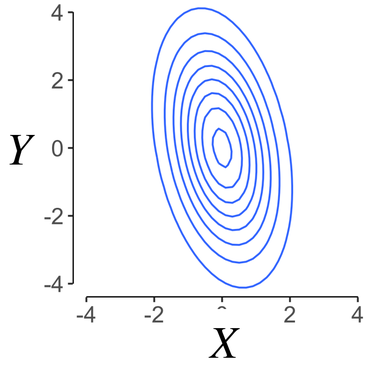
$\rho = 0, \sigma_X = \sigma_Y = 2$



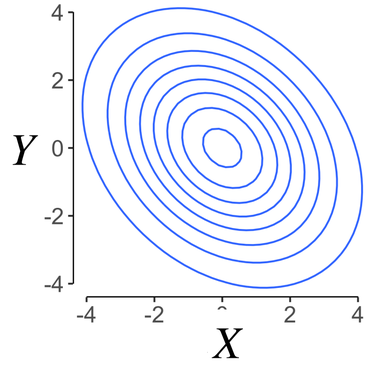
$\rho = -0.3, \sigma_X = \sigma_Y = 1$



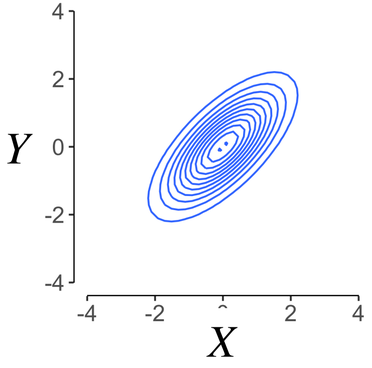
$\rho = -0.3, \sigma_X = 1, \sigma_Y = 2$



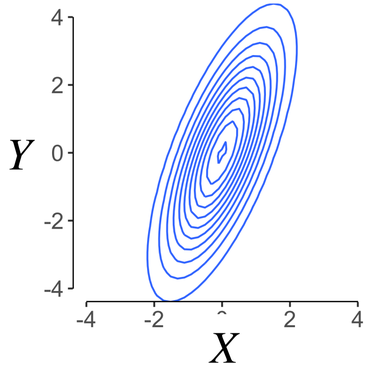
$\rho = -0.3, \sigma_X = \sigma_Y = 2$



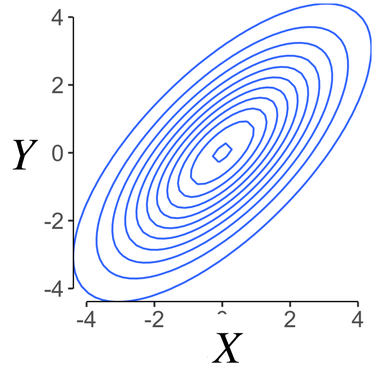
$\rho = 0.7, \sigma_X = \sigma_Y = 1$



$\rho = 0.7, \sigma_X = 1, \sigma_Y = 2$



$\rho = 0.7, \sigma_X = \sigma_Y = 2$



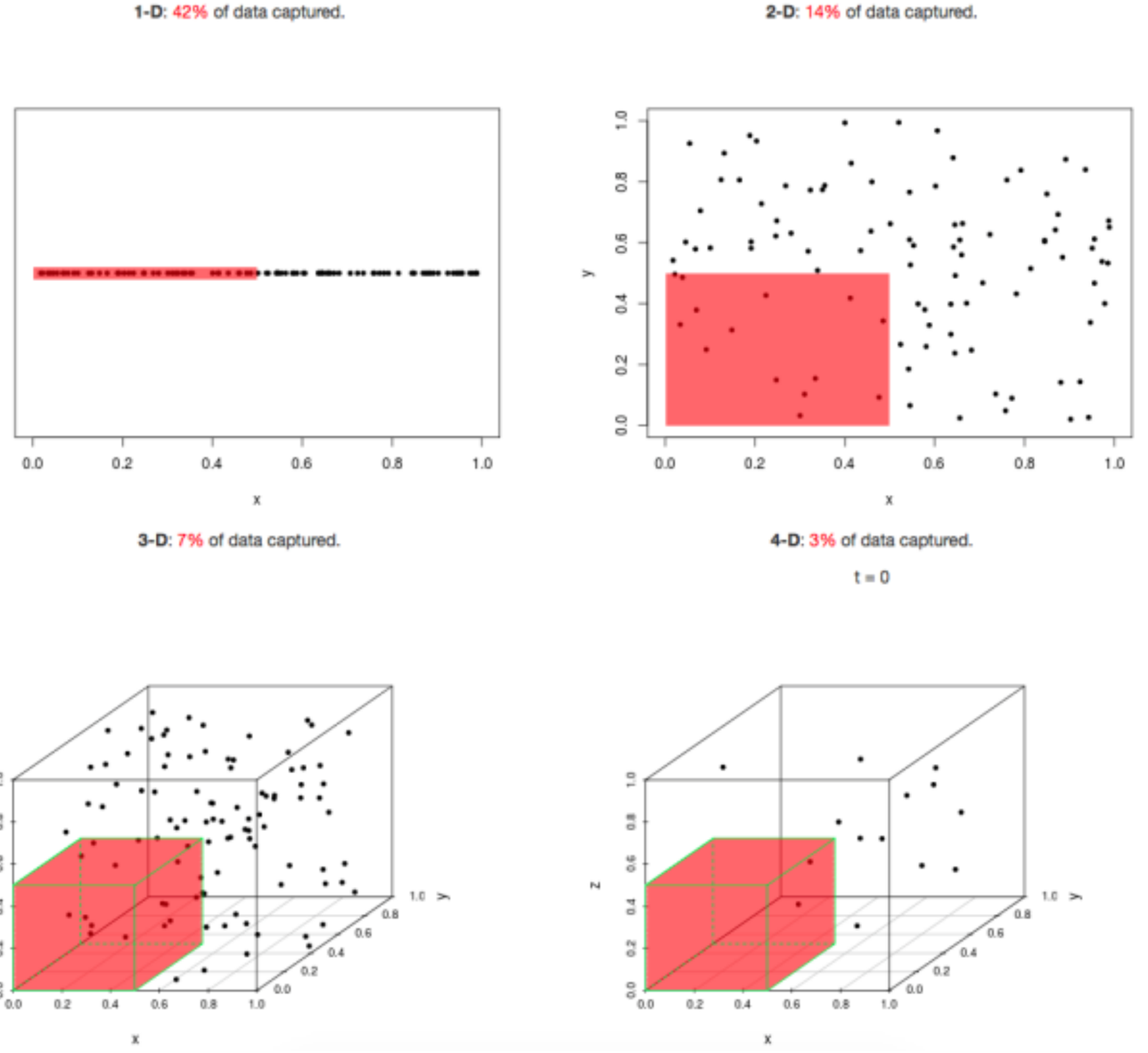
$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

Axis-aligned: $\rho = 0$

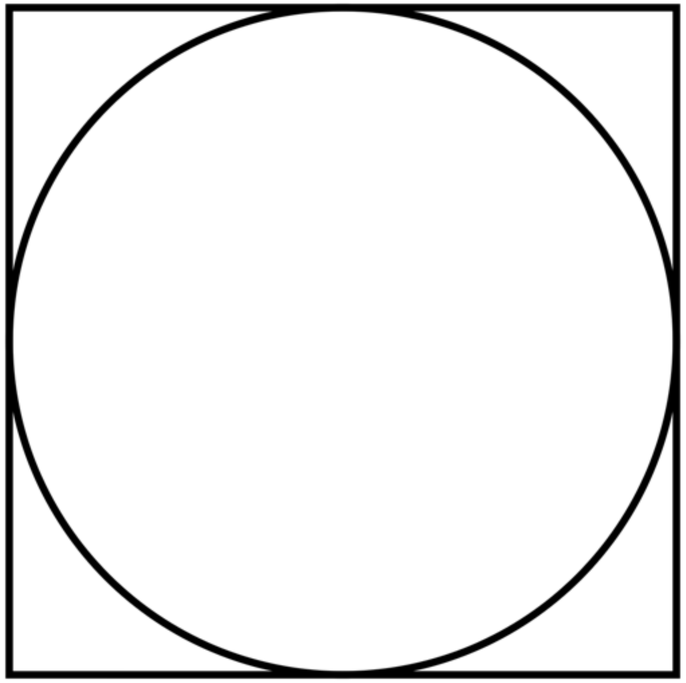
Spherical: all σs are the same

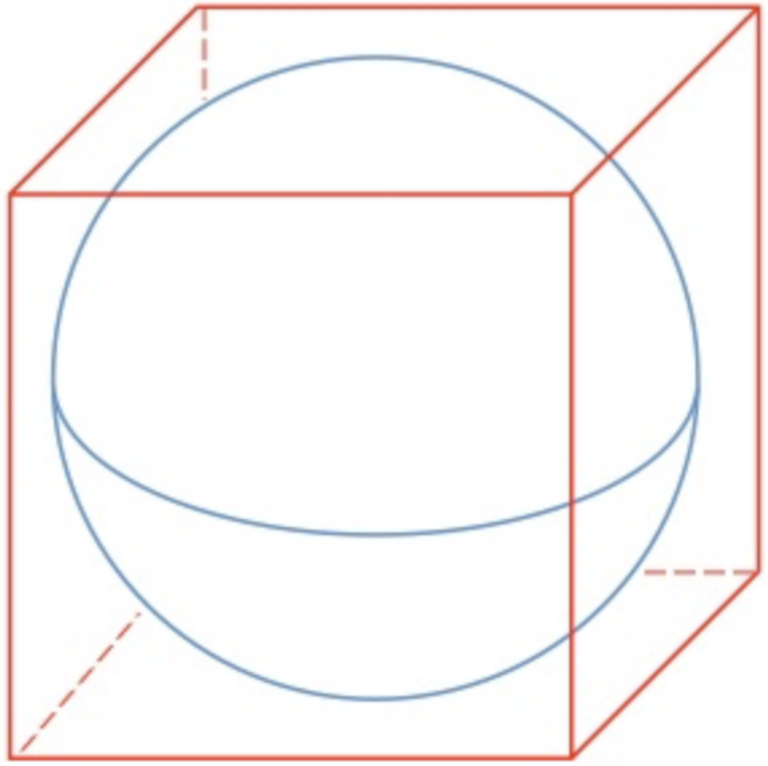
The Curse of Dimensionality

- Nearest neighbor breaks down in high-dimensional spaces because the “neighborhood” becomes very large.
- Suppose we have 5000 points uniformly distributed in the unit hypercube and we want to apply the 5-nearest neighbor algorithm.
- Suppose our query point is at the origin.
  - 1D –
    - On a one dimensional line, we must go a distance of 5/5000 = 0.001 on average to capture the 5 nearest neighbors
  - 2D –
    - In two dimensions, we must go sqrt(0.001) to get a square that contains 0.001 of the volume
  - D –
    - In D dimensions, we must go $(0.001)^{1/D}$



Slides from David Sonntag, NYU





One way to illustrate the "vastness" of high-dimensional Euclidean space is to compare the proportion of an inscribed hypersphere with radius r and dimension d, to that of a hypercube with edges of length $2r$. The volume of such a sphere is $\frac{2r^d \pi^{d/2}}{d\,\Gamma(d/2)}$, where Γ is the gamma function, while the volume of the cube is $(2r)^d$. As the dimension d of the space increases, the hypersphere becomes an insignificant volume relative to that of the hypercube. This can clearly be seen by comparing the proportions as the dimension d goes to infinity:

$$\frac{V_{\text{hypersphere}}}{V_{\text{hypercube}}} = \frac{\pi^{d/2}}{d2^{d-1}\Gamma(d/2)} \to 0 \text{ as } d \to \infty.$$

Higher order statistics are more vulnerable

To the curse of dimensionality

- Empirical covariance tends to have troubles before empirical means
- Both kinds of MLE estimates can be regularized through “shrinkage”
- $\mathbf{m}_{\text{shrunk}} = (1 - \alpha)\mathbf{m}_{\text{MLE}} + \alpha C\mathbf{I}$ where α is a hyper parameter that mixes the MLE with an uninformative prior (identity matrix).
- For example uses see https://scikit-learn.org/stable/modules/covariance.html#shrunk-covariance
-