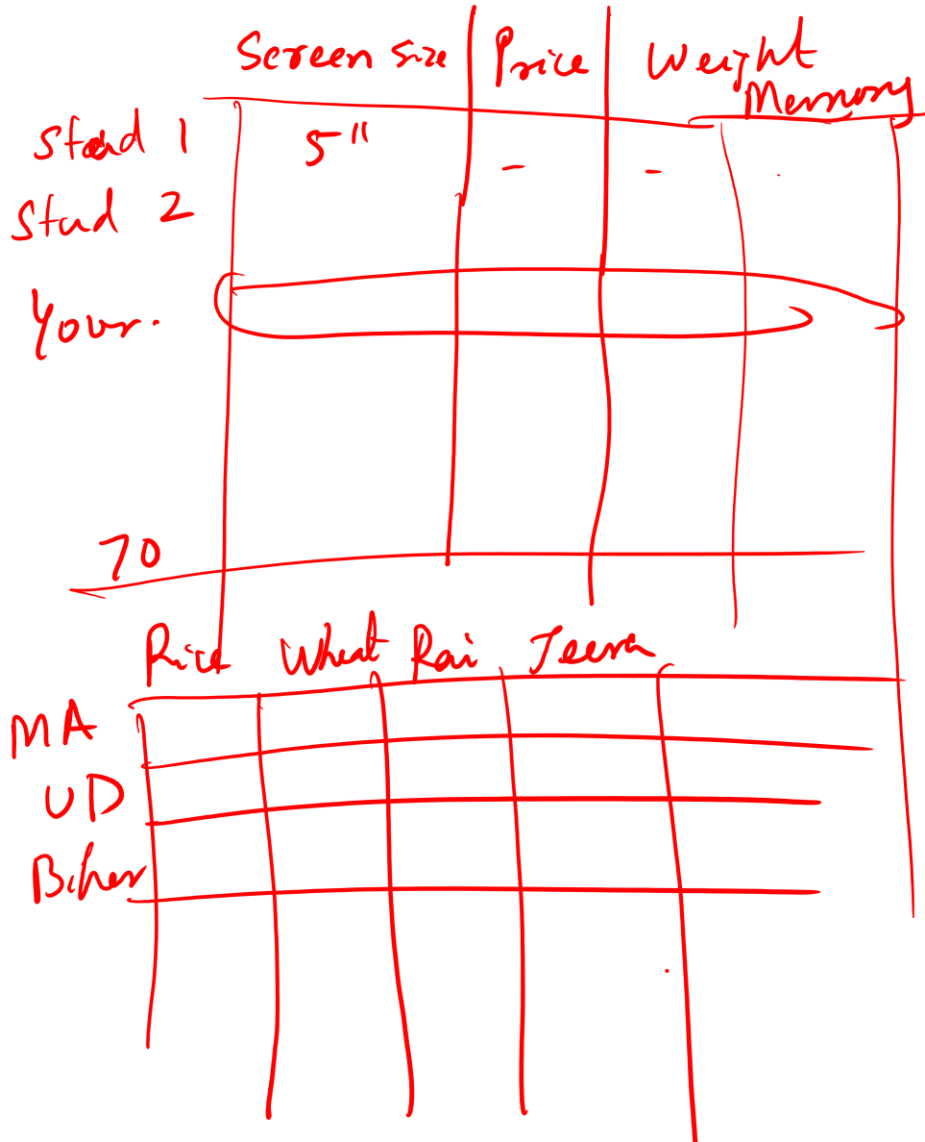
Distance between two points

Two points: $\mathrm{P} = [x_1, x_2, \ldots, x_p]'$, $Q = [y_1, y_2, \ldots, y_p]'$

Euclidean distance between them:

$$d(P,Q) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_p - y_p)^2}$$

- Gives equal importance to all dimensions.
- Does not make sense always.
  - Example mobiles phone statistics
    - x_1, x_2 refers to screen size in cms, and cost in rupees
    - A 2 point difference in screen size is more significant than a 2 rupee difference in cost.
  - Example: states with production of various grains
    - x_1, x_2 refers to production of rice Vs mustard in kilo tons.



Mahalonobis distance

Prasanta Chandra Mahalanobis

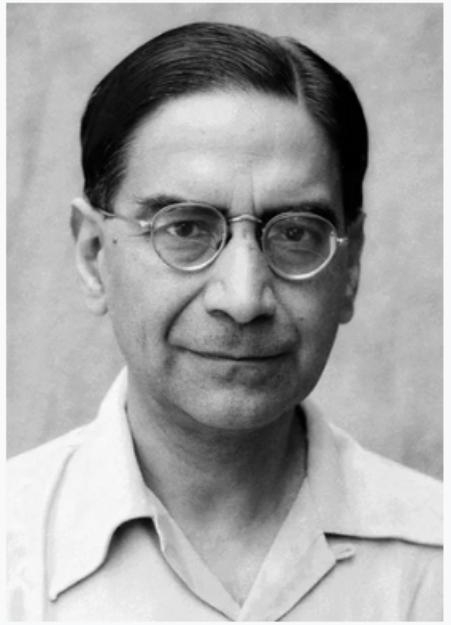
25 languages

Article Talk Read Edit View history Tools

From Wikipedia, the free encyclopedia

Prasanta Chandra Mahalanobis OBE, FNA,[5] FASc,[6] FRS[2] (29 June 1893– 28 June 1972) was an Indian scientist and statistician. He is best remembered for the Mahalanobis distance, a statistical measure, and for being one of the members of the first Planning Commission of free India. He made pioneering studies in anthropometry in India. He founded the Indian Statistical Institute, and contributed to the design of large-scale sample surveys.[2][7][4][8] For his contributions, Mahalanobis has been considered the Father of statistics in India.[9]

Prasanta Chandra Mahalanobis



Born	29 June 1893 Calcutta, Bengal Presidency, British India
Died	28 June 1972 (aged 78) Kolkata, West Bengal, India
Alma mater	University of Calcutta (BSc) King's College, Cambridge (BA)[2]

Early life [edit]

Young Mahalanobis

Mahalanobis was born on 29 June 1893, in Calcutta, Bengal Presidency (now West Bengal). Mahalanobis belonged to a prominent Bengali Brahmin family of landed gentry in Bikrampur, Dhaka, Bengal Presidency (now in Bangladesh).[10][11] His grandfather Gurucharan (1833–1916) moved to Calcutta in 1854 and built up a business, starting a chemist shop in 1860. Gurucharan was influenced by Debendranath Tagore (1817–1905), father of the Nobel Prize-winning poet, Rabindranath Tagore. Gurucharan

Mahalonobis/Statistical distance between two points

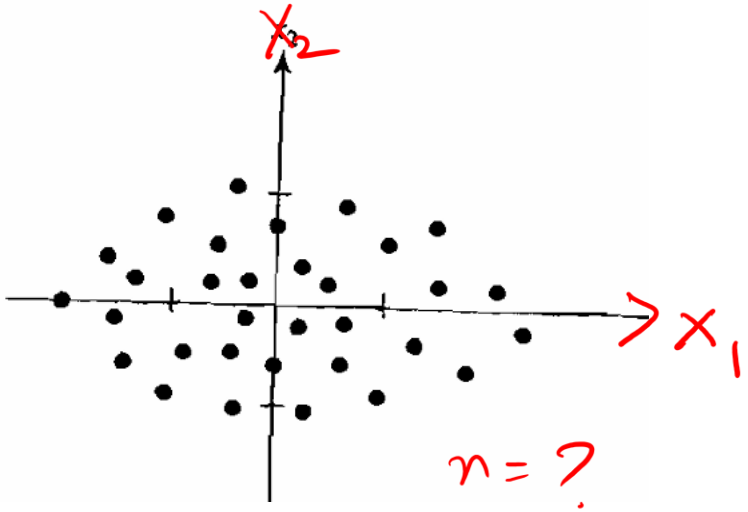
Two points: $\mathrm{P} = [x_1, x_2, \ldots, x_p]'$, $Q = [y_1, y_2, \ldots, y_p]'$

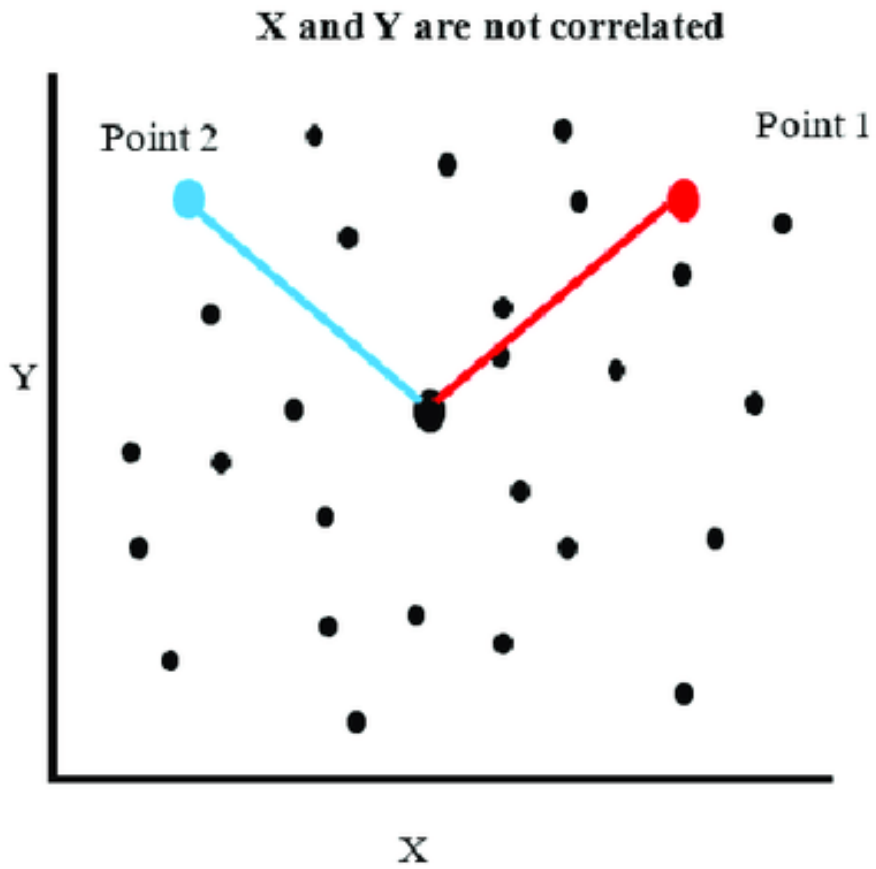
- If points are more spread out in one dimension, then we expect distance between any two random points to be larger in that dimension.

- A new distance function: ~~standard deviation~~ variance scaled.

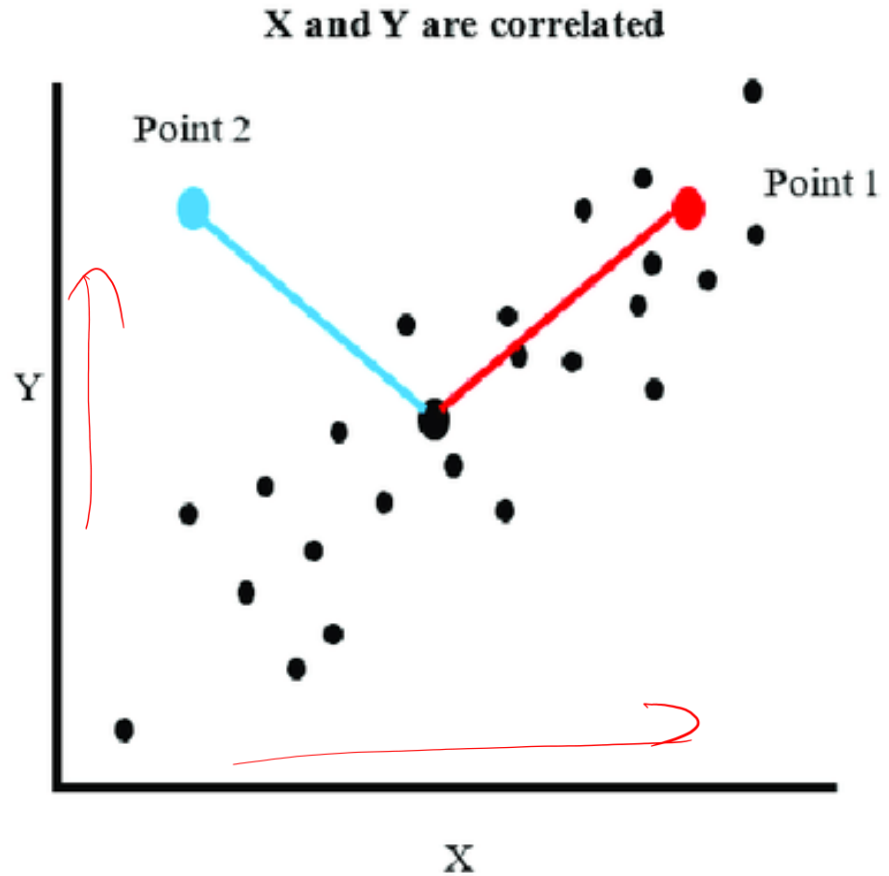
$$d(P,Q) = \sqrt{\frac{(x_1 - y_1)^2}{s_{11}} + \frac{(x_2 - y_2)^2}{s_{22}}}$$

variance along dimension 1





When X and Y are not correlated, the Euclidean distance from the Centroid can be useful to inter if a point is member of the distribution



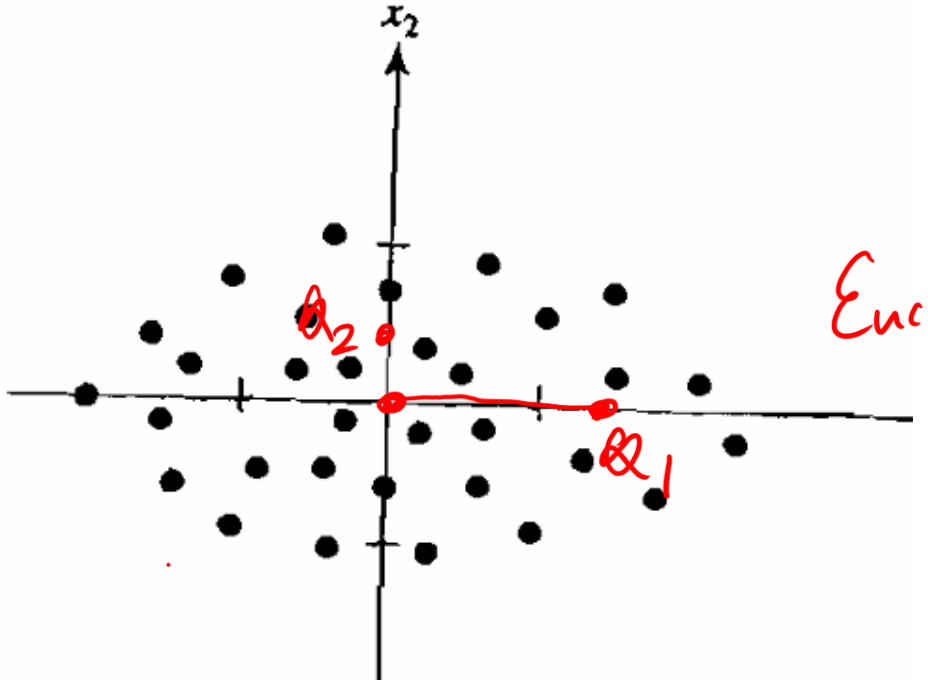
Point one and two have the same Euclidean Distance from Centroid but only point one is a member of the distribution. to detect point two as outlier, dist. (point two, centroid) should be much higher than dist. (point one, Centroid) Mahalanobis distance can be used here instead.

Variance scaled distance:

- Distance to origin

$$d(P \equiv O;\ Q) \equiv \sqrt{\frac{y_1^2}{s_{11}} + \frac{y_2^2}{s_{22}} + \cdots \frac{y_p^2}{s_{pp}}}$$

- Points closer by Euclidean distance might get further by variance-scaled distance



$Euc\,D(0, Q_2) < Euc\,D(0, Q_1)$

but $d_D(0, Q_2) > d_D(0, Q_1)$

Contours of equal distance from the origin

Set of all points at the same distance from the mean (= origin here)

P = 2 ;

$$\frac{x_1^2}{s_{11}} + \frac{x_2^2}{s_{22}} = c$$

Constant

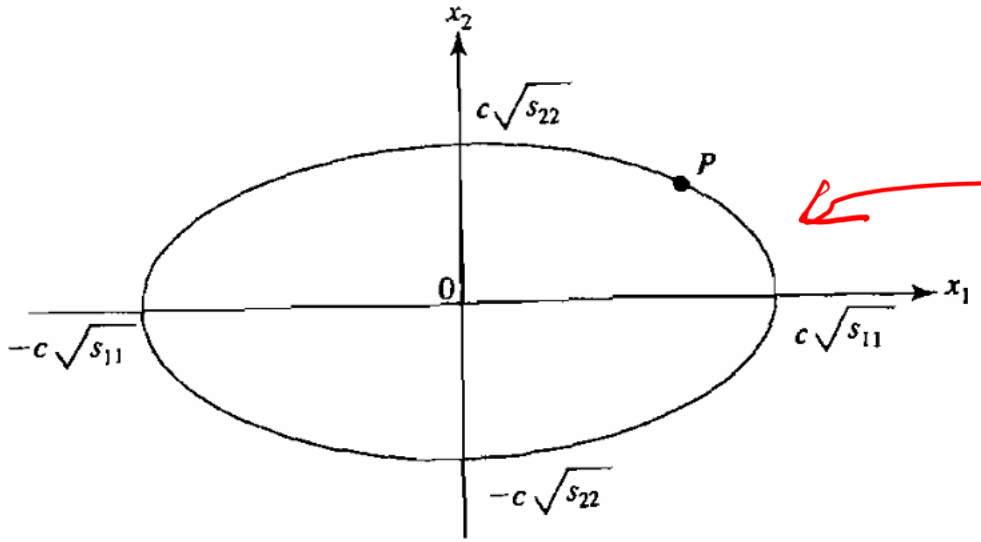
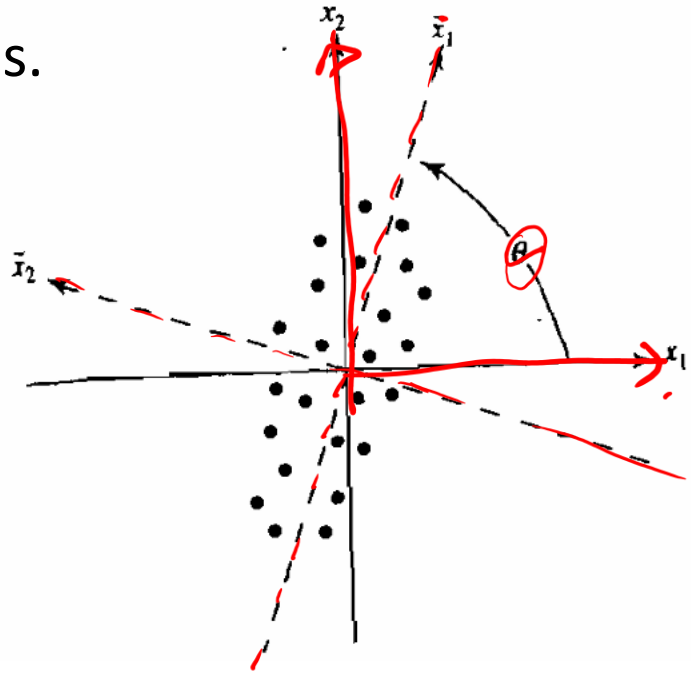


Figure 1.21 The ellipse of constant statistical distance $d^2(O, P) = x_1^2/s_{11} + x_2^2/s_{22} = c^2$.

Distance when data has correlation among variables

Define a new co-ordinate system where the correlation vanishes.
We will see how to do that in general.



$$\tilde{x}_1 = x_1 \cos(\theta) + x_2 \sin(\theta)$$
$$\tilde{x}_2 = -x_1 \sin(\theta) + x_2 \cos(\theta)$$

$$d(O,P) = \sqrt{\frac{\tilde{x}_1^2}{\tilde{s}_{11}} + \frac{\tilde{x}_2^2}{\tilde{s}_{22}}}$$

$$d(O,P) = \sqrt{a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2}$$

a_{11}, a_{12}, a_{22} are functions of s_{11}, s_{12}, s_{22}
Exact form will be provided later.

$$d(P,Q) = \sqrt{a_{11}(x_1 - y_1)^2 + 2a_{12}(x_1 - y_1)(x_2 - y_2) + a_{22}(x_2 - y_2)^2}$$

Generalization to multiple dimensions:

$$d^2(P,Q) = \sum_{i=1}^{p} \sum_{j=1}^{p} a_{ij}(x_i - y_i)(x_j - y_j)$$

$$A_{p \times p} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ \vdots & & & \vdots \\ a_{p1} & & & a_{pp} \end{bmatrix} = S^{-1}$$

Statistical distance between points in matrix notation

- Square of distance between two points

$$(x-y)^T A (x-y) \qquad \text{eg: } p=2$$

$$\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}\right) = \begin{bmatrix} x_1-y_1 \\ x_2-y_2 \end{bmatrix}$$

$$\begin{bmatrix} x_1-y_1 & x_2-y_2 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} \begin{bmatrix} x_1-y_1 \\ x_2-y_2 \end{bmatrix}$$

$$= (x_1-y_1)^2 a_{11} + 2a_{12}(x_1-y_1)(x_2-y_2) + a_{22}(x_2-y_2)^2$$

- Distance to origin $d^2(0,P) \qquad P = \begin{bmatrix} x_1 \\ x_p \end{bmatrix}$

$$x^T A x$$

Valid A are those where distance is always non-negative. $x^T A x \geq 0$

Quadratic forms

- Given a vector x of size p, and a square matrix A of size p x p, the quadratic form of x is

$$Q_A(x) = x^T A x$$

$$Q(\mathbf{x}) = [x_1 \quad x_2]\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}\begin{bmatrix}x_1\\ x_2\end{bmatrix} = x_1^2 + 2x_1x_2 + x_2^2$$

$$Q(\mathbf{x}) = [x_1 \quad x_2 \quad x_3]\begin{bmatrix}1 & 3 & 0\\ 3 & -1 & -2\\ 0 & -2 & 2\end{bmatrix}\begin{bmatrix}x_1\\ x_2\\ x_3\end{bmatrix} = x_1^2 + 6x_1x_2 - x_2^2 - 4x_2x_3 + 2x_3^2$$

Positive definite and semi-definite matrices

- A square matrix A is positive definite if for all vectors y, the value of the quadratic form is > 0. $y^T A y > 0 \quad \forall y$
  - Positive semi-definite if quadratic form is >= 0.

$y^T A y \geq 0 \quad \forall y.$

Covariance matrix is positive semi-definite

$$S = \left(X_{n\times p} - 1_{n\times 1}\bar{x}\right)^T \left(X_{n\times p} - 1_{n\times 1}\bar{x}\right)$$

$$= \tilde{X}^T \tilde{X}$$

$$\forall y \quad y^T S y \geq 0$$

$$y^T \tilde{X}^T \tilde{X} y = \left(\tilde{X} y\right)^T \tilde{X} y = U^T U \geq 0$$

- Inverse of a positive semi-definite matrix is positive semi-definite [Proof in terms of spectral decomposition will be shown later.]

Population mean and covariance

- Let $f(X_1, \ldots, X_p)$ be joint distribution over p variable.

$$\mu_i = \begin{cases} \int_{-\infty}^{\infty} x_i f_i(x_i)\, dx_i \\ \sum_{\text{all } x_i} x_i p_i(x_i) \end{cases}$$

$$\sigma_{ik} = E(X_i - \mu_i)(X_k - \mu_k)$$

$$= \begin{cases} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_i)(x_k - \mu_k) f_{ik}(x_i, x_k) dx_i\, dx_k \\ \sum_{\text{all } x_i} \sum_{\text{all } x_k} (x_i - \mu_i)(x_k - \mu_k) p_{ik}(x_i, x_k) \end{cases}$$

In matrix notation

$$E(\mathbf{X}) = \begin{bmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_p) \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}$$

Warning! The X here refers to a random vector of length p. This should not be confused with the X used to denote the data matrix in earlier slides which is of size n x p

Covariance

$$\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix}$$

$$\boldsymbol{\Sigma} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})'$$

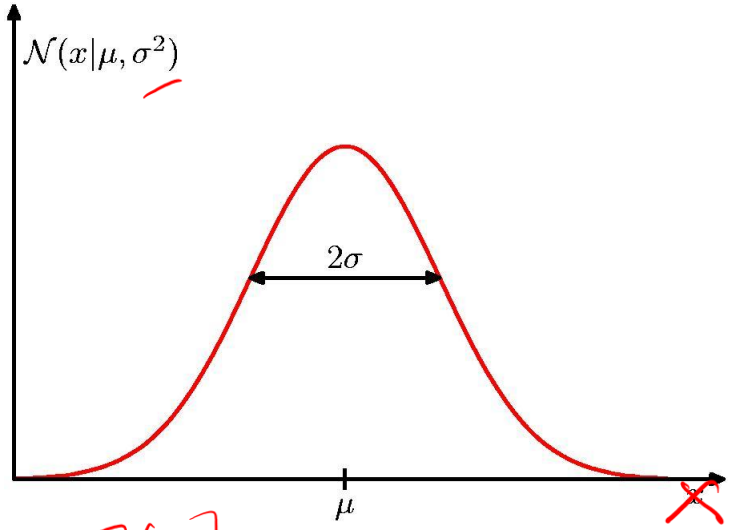
$$= E\left(\begin{bmatrix} X_1 - \mu_1 \\ X_2 - \mu_2 \\ \vdots \\ X_p - \mu_p \end{bmatrix} [X_1 - \mu_1, X_2 - \mu_2, \ldots, X_p - \mu_p] \right)$$

$$= E\begin{bmatrix} (X_1 - \mu_1)^2 & (X_1 - \mu_1)(X_2 - \mu_2) & \cdots & (X_1 - \mu_1)(X_p - \mu_p) \\ (X_2 - \mu_2)(X_1 - \mu_1) & (X_2 - \mu_2)^2 & \cdots & (X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (X_p - \mu_p)(X_1 - \mu_1) & (X_p - \mu_p)(X_2 - \mu_2) & \cdots & (X_p - \mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} E(X_1 - \mu_1)^2 & E(X_1 - \mu_1)(X_2 - \mu_2) & \cdots & E(X_1 - \mu_1)(X_p - \mu_p) \\ E(X_2 - \mu_2)(X_1 - \mu_1) & E(X_2 - \mu_2)^2 & \cdots & E(X_2 - \mu_2)(X_p - \mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ E(X_p - \mu_p)(X_1 - \mu_1) & E(X_p - \mu_p)(X_2 - \mu_2) & \cdots & E(X_p - \mu_p)^2 \end{bmatrix}$$

Multidimensional Gaussian Distribution

The Gaussian Distribution



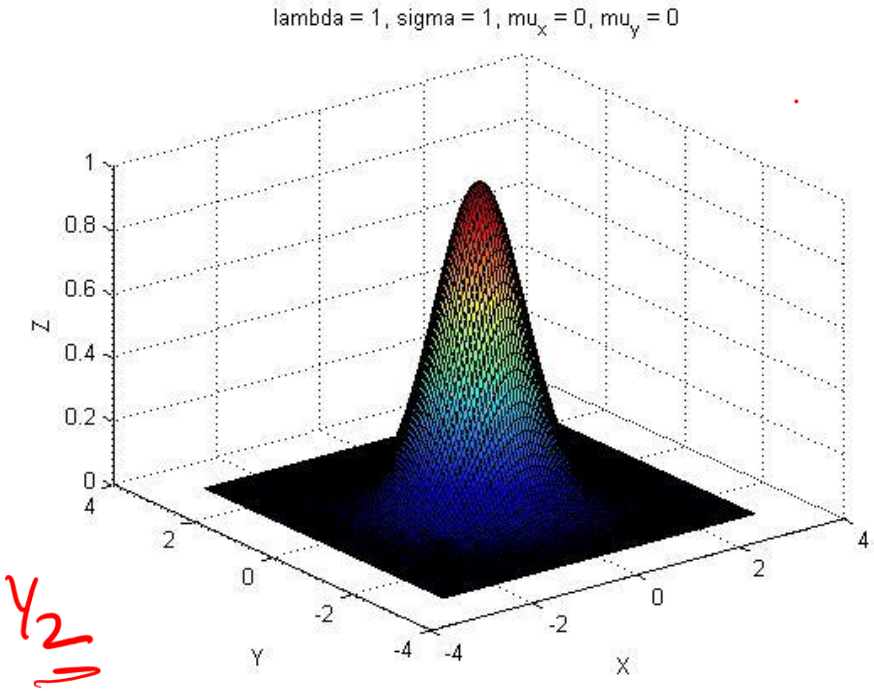
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

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

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_p \end{bmatrix} \qquad \Sigma_{p\times p} = \begin{bmatrix} \sigma_{11} & \cdots \\ \vdots & \ddots \end{bmatrix}$$

$$|\Sigma|^{1/2}$$



Bi-variate Gaussian density

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

$$\rho_{12} = \frac{\sigma_{12}}{\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}}$$

is

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2} \begin{bmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11} \end{bmatrix}$$

Introducing the correlation coefficient ρ_{12} by writing $\sigma_{12} = \rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}$, we obtain $\sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$, and the squared distance becomes

$$(\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$$

$$= [x_1 - \mu_1, x_2 - \mu_2] \frac{1}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)} \begin{bmatrix} \sigma_{22} & -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} \\ -\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}} & \sigma_{11} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}$$

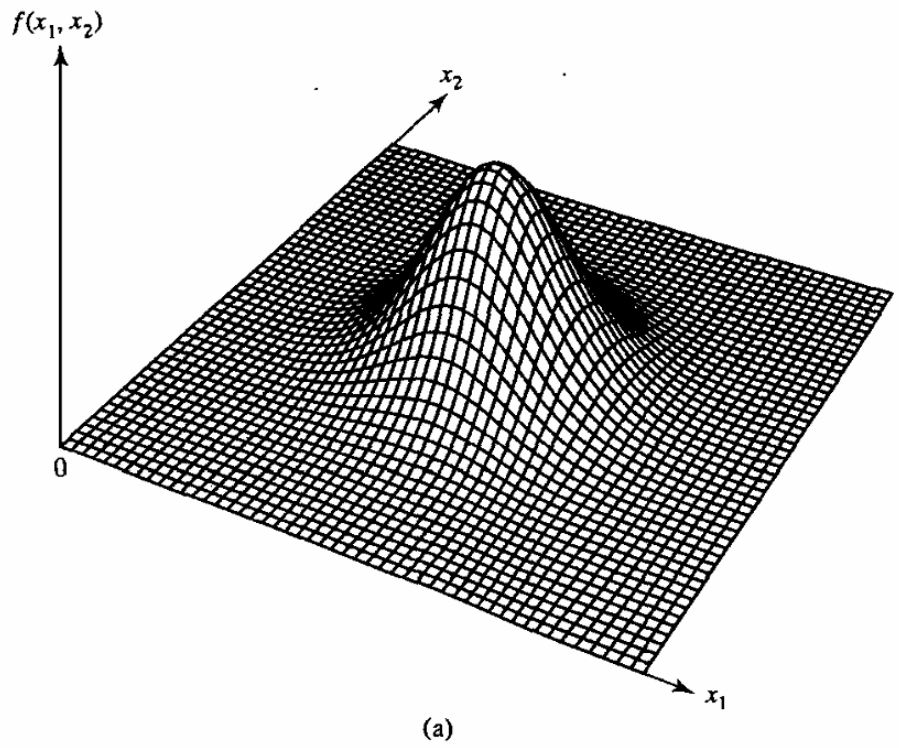
$$= \frac{\sigma_{22}(x_1 - \mu_1)^2 + \sigma_{11}(x_2 - \mu_2)^2 - 2\rho_{12}\sqrt{\sigma_{11}}\sqrt{\sigma_{22}}(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}$$

$$= \frac{1}{1 - \rho_{12}^2}\left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)\right] \quad (4\text{-}5)$$

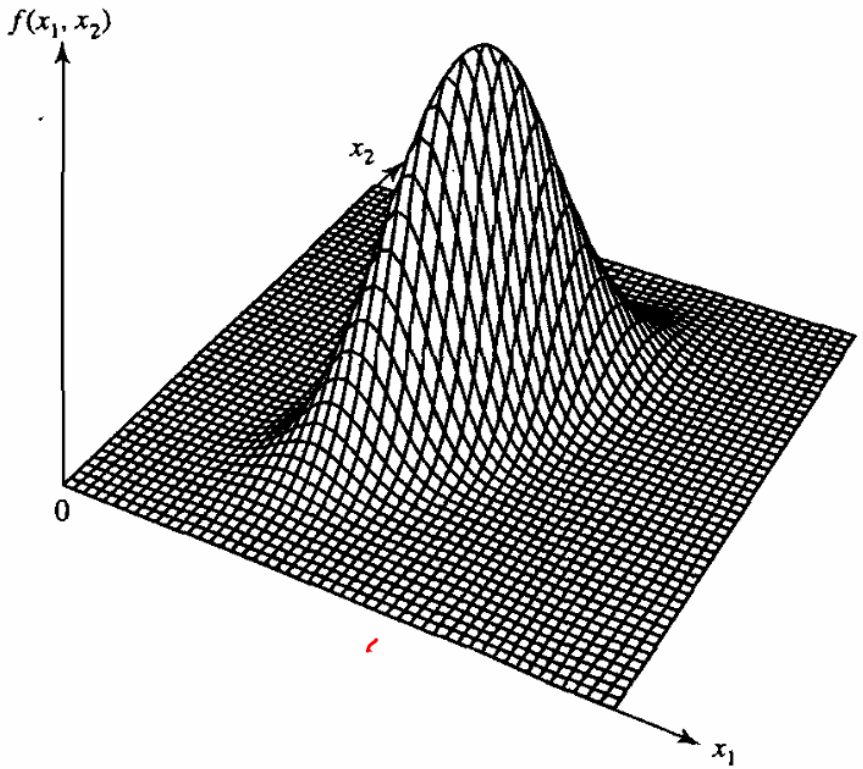
$$|\boldsymbol{\Sigma}| = \sigma_{11}\sigma_{22} - \sigma_{12}^2 = \sigma_{11}\sigma_{22}(1 - \rho_{12}^2)$$

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1 - \rho_{12}^2)}} \times \exp\left\{-\frac{1}{2(1 - \rho_{12}^2)}\left[\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)^2 + \left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)^2 - 2\rho_{12}\left(\frac{x_1 - \mu_1}{\sqrt{\sigma_{11}}}\right)\left(\frac{x_2 - \mu_2}{\sqrt{\sigma_{22}}}\right)\right]\right\} \tag{4-6}$$

Visualization



(a)



(b)

Figure 4.2 Two bivariate normal distributions. (a) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = 0$. (b) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = .75$.

Constant density or contour plots

$$\textit{Constant probability density contour} = \{\text{all } \mathbf{x} \text{ such that } (\mathbf{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = c^2\}$$

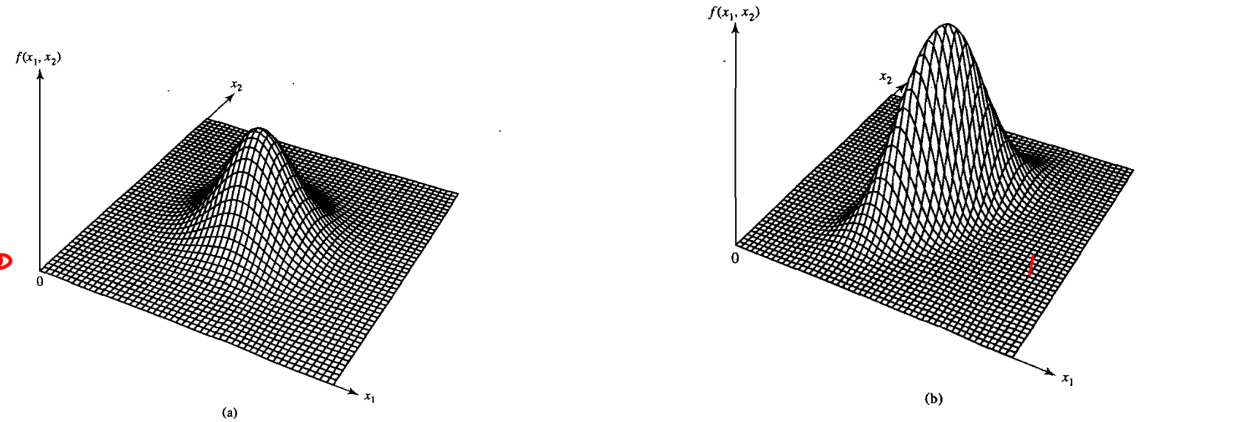


Figure 4.2 Two bivariate normal distributions. (a) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = 0$. (b) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = .75$.

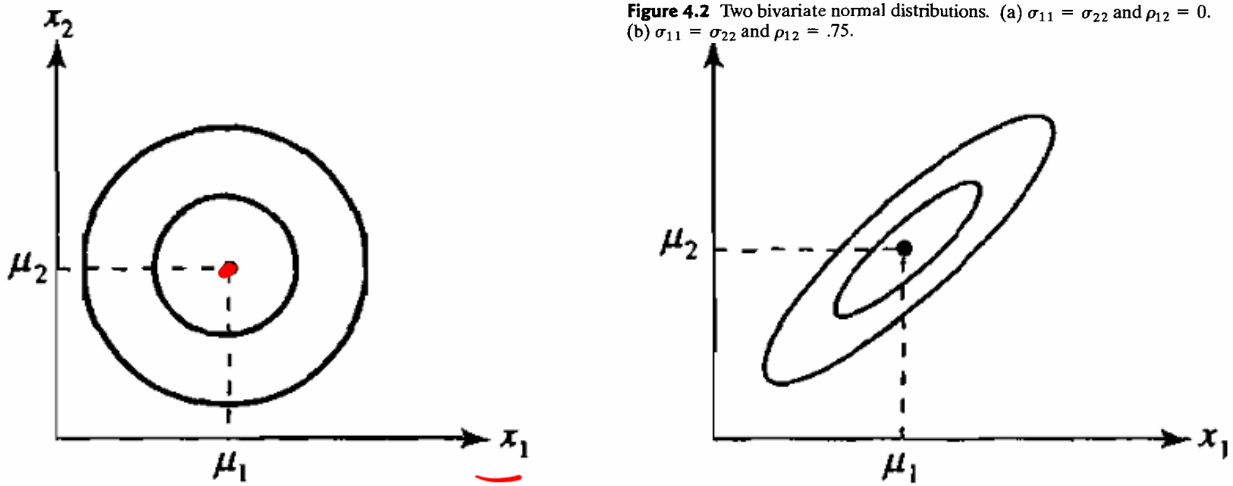
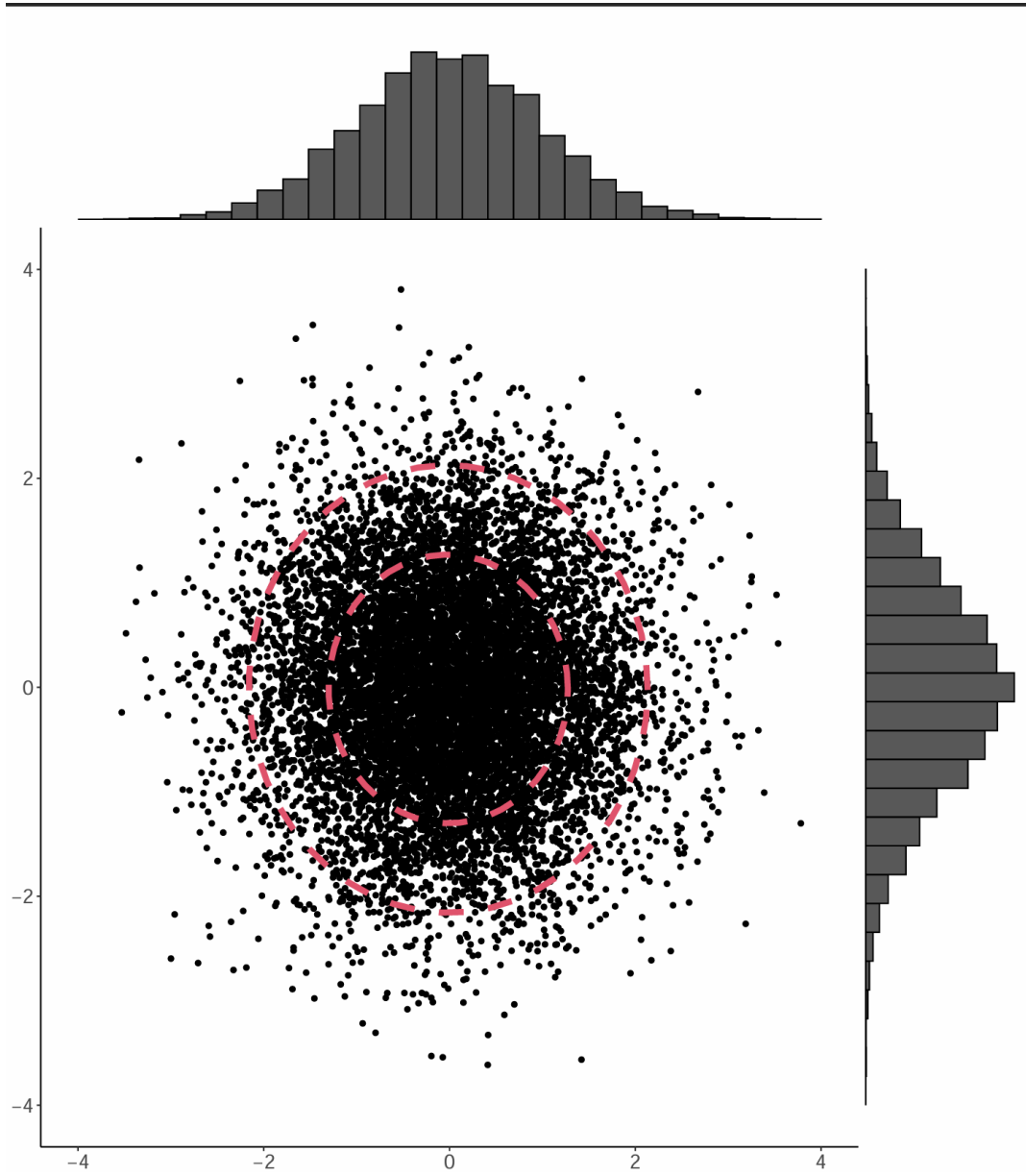


Figure 4.4 The 50% and 90% contours for the bivariate normal distributions in Figure 4.2.

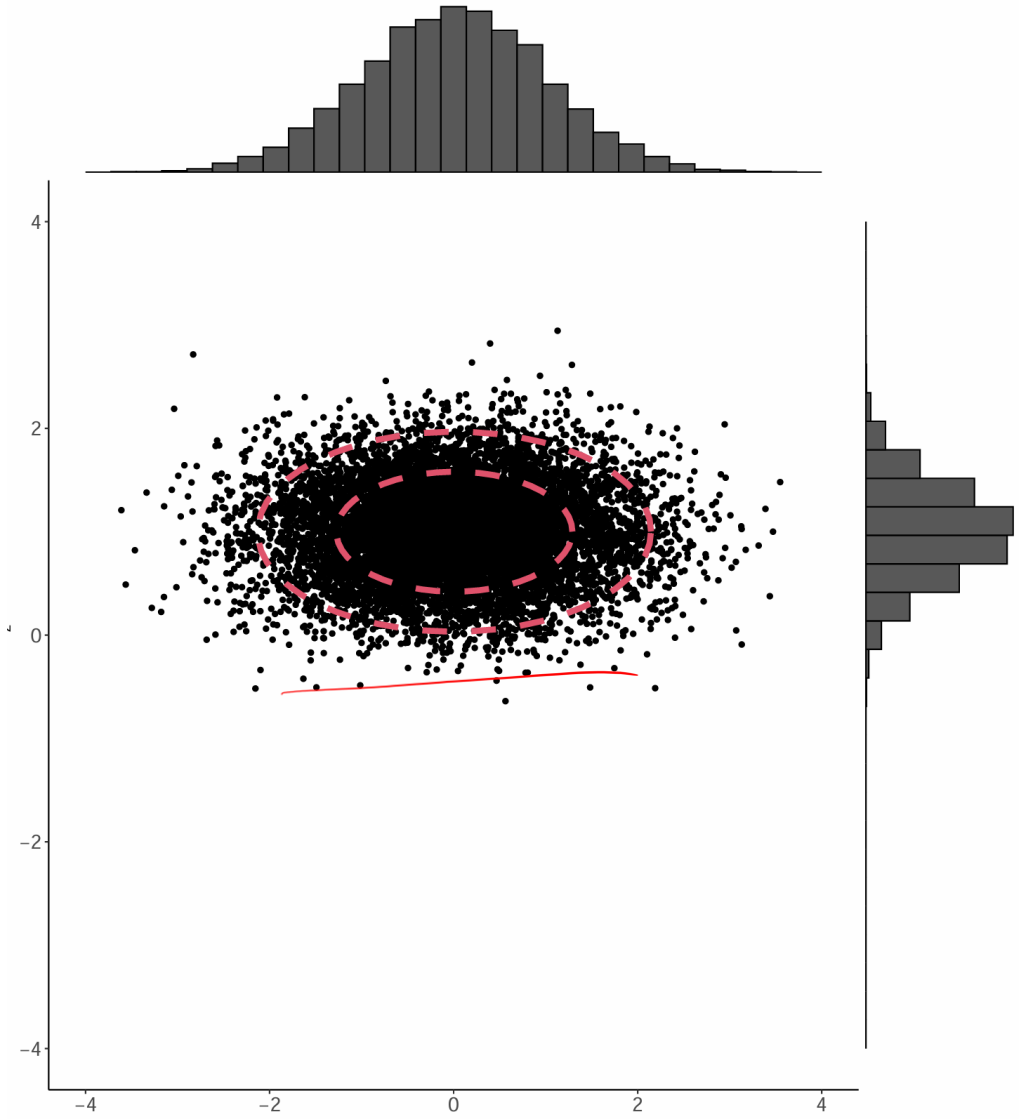
Visualizing in 2-D via contours



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

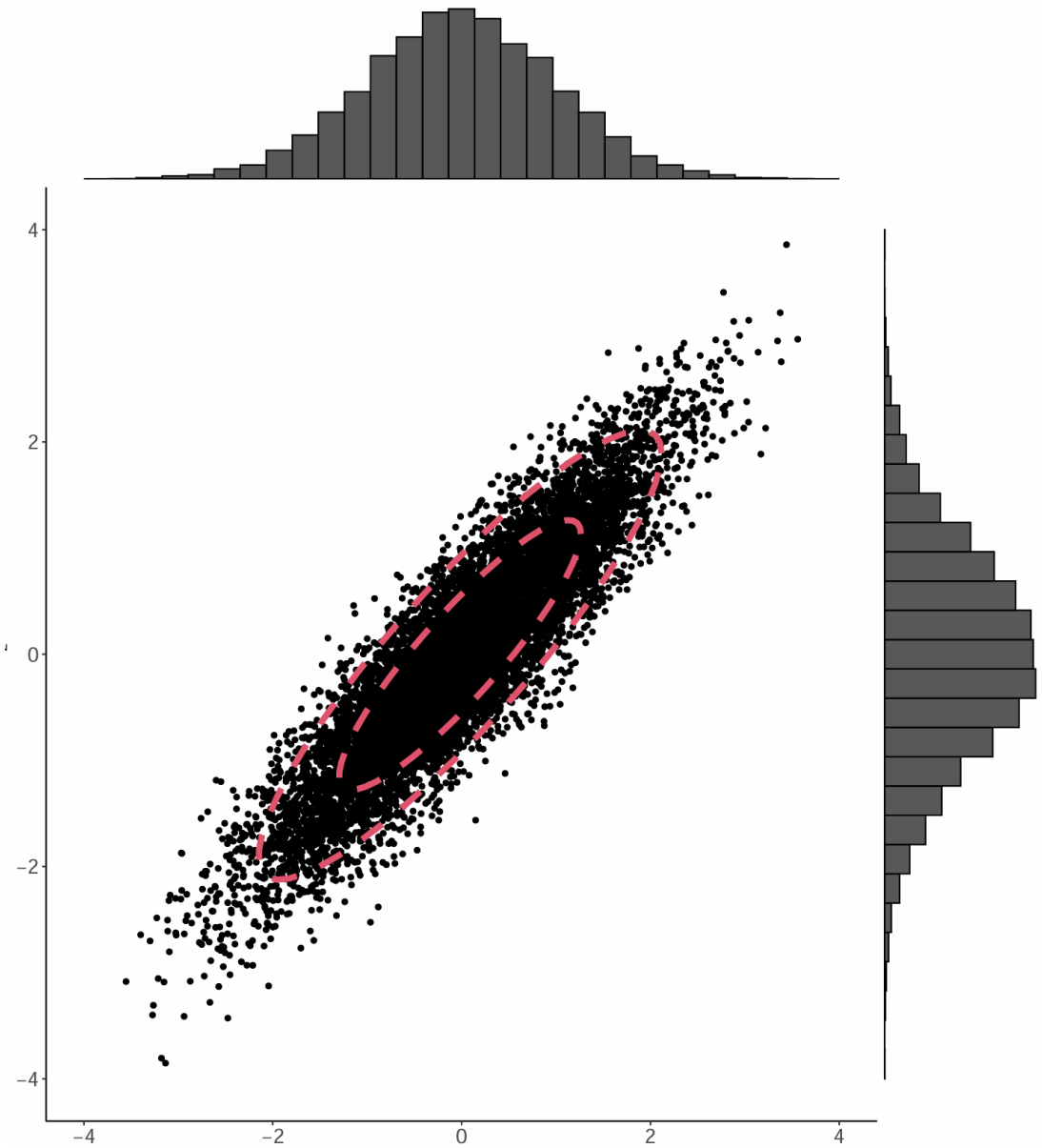
Different variance along each dimension



$$\mu = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 0.2 \end{pmatrix}$$

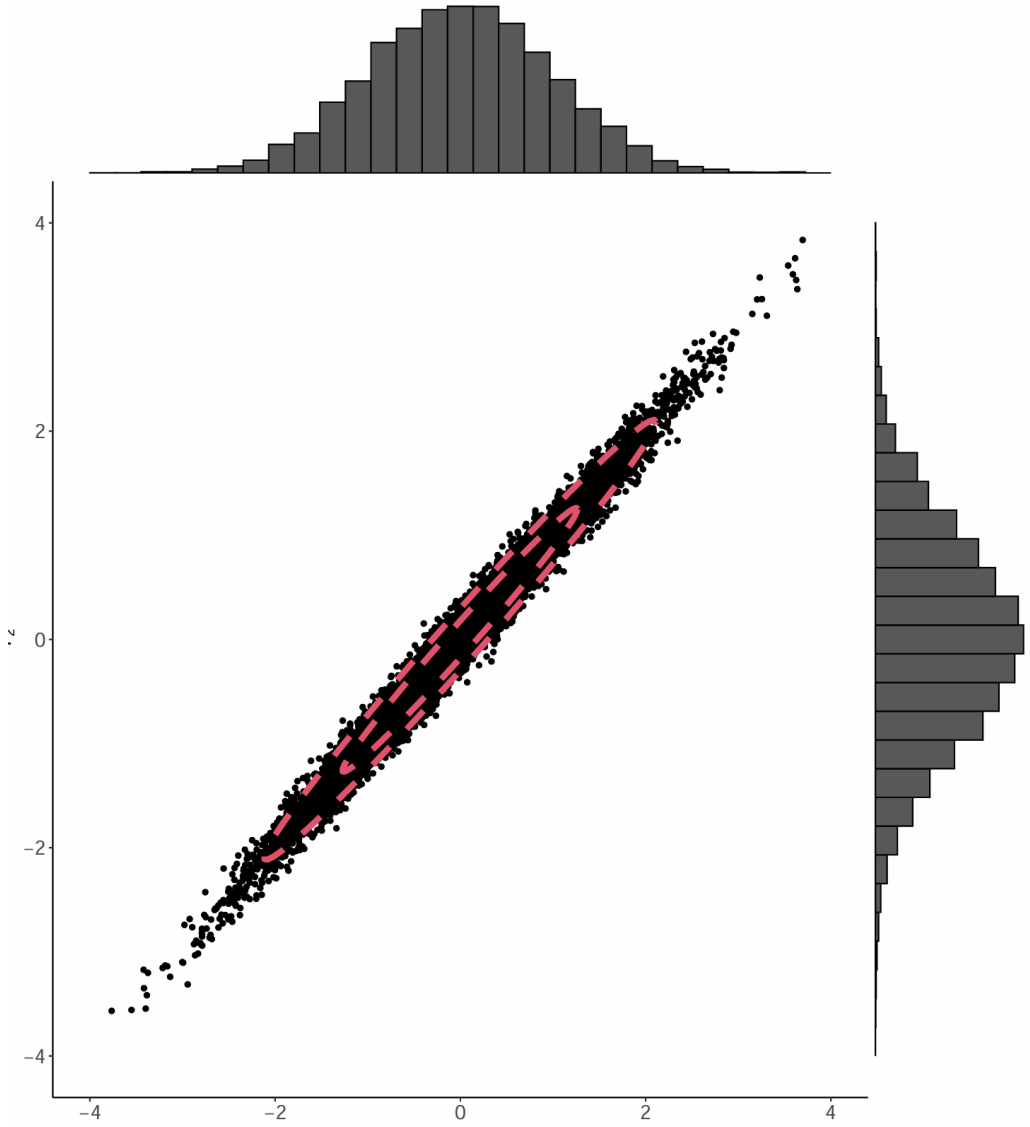
Correlated variables



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$$

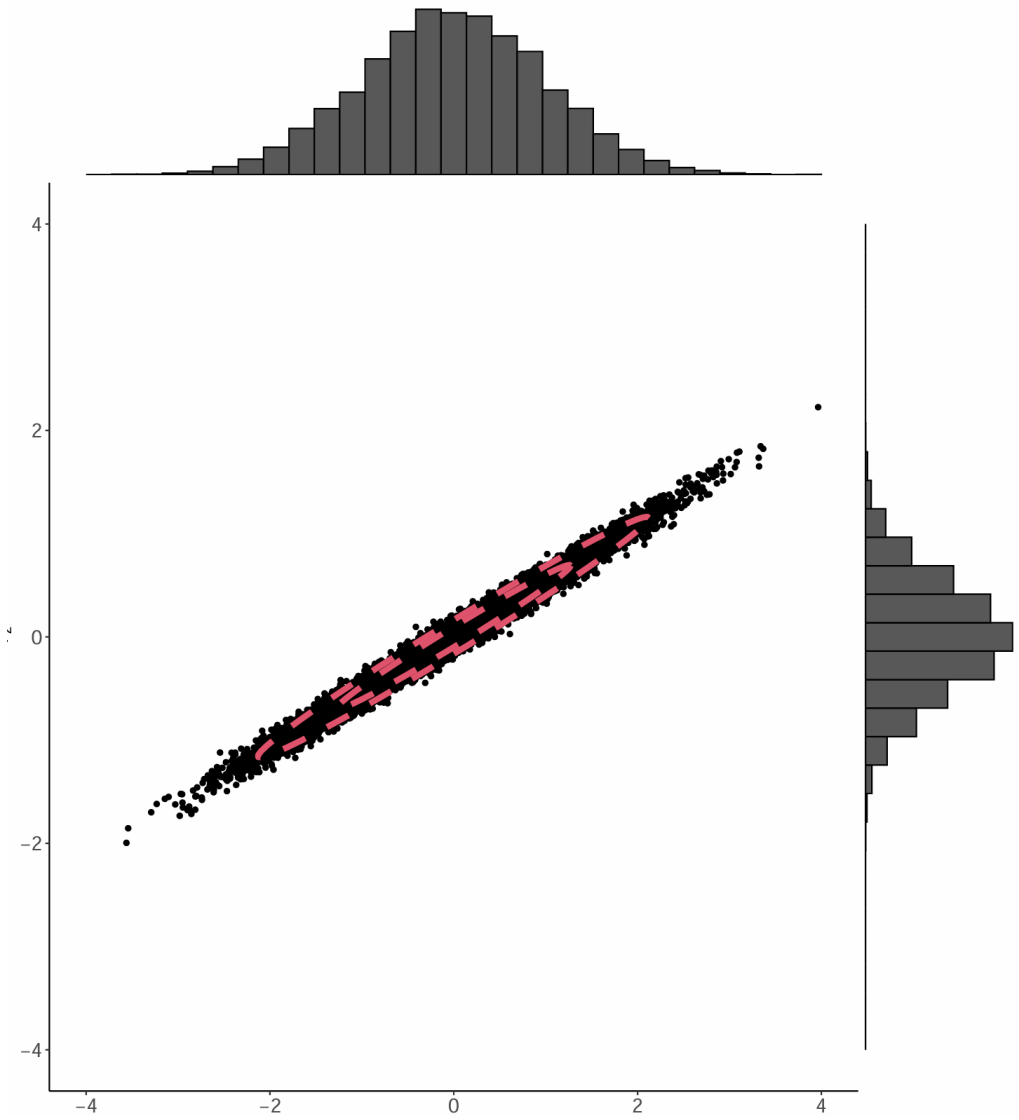
Highly correlated variables



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}$$

Correlation with different variance.



$$\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} 1 & 0.54 \\ 0.54 & 0.3 \end{pmatrix}$$

$Cor(Y_1, Y_2) = 0.54/\sqrt{(0.3)} = 0.99$