University of Southampton

SESA6085 – Advanced Aerospace Engineering Management

Lecture 4

2023-2024

Dr David Toal

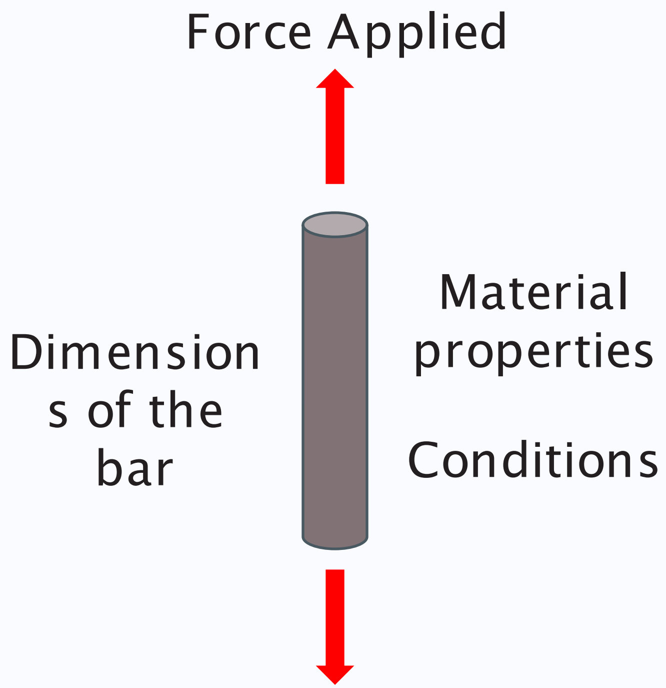
Multi-variate Distributions

Continuous PDFs

- Previously we have considered a number of PDFs
  - Gaussian, log-normal, exponential, Weibull etc.
- Each of which operates along a single variable
- Is this realistic?
- In some cases yes, in other cases no

Single or Multiple Variables?

- Let us consider a simple example
- A metal bar is subject to tension what factors will impact its failure?



Multivariate Models

- Clearly in this case there are a number of factors which could impact the failure and hence reliability
- These factors are not independent
- Multivariate models are used to help model such dependencies and can come in a variety of forms e.g.
  - Normal, log-normal, exponential, Weibull etc.
- In a lot of cases there is no closed form solution to determine their parameters

Univariate Models

- Recall previously that our univariate models were the result of single observations, x, from a series of identical experiments
- From this data we could define the probability of an observation occurring

$$F(x) = P(X \leq x)$$

- Similarly, we could calculate the probability of an observation occurring between two bounds a & b as:

$$P(a < X \leq b) = F(b) - F(a)$$

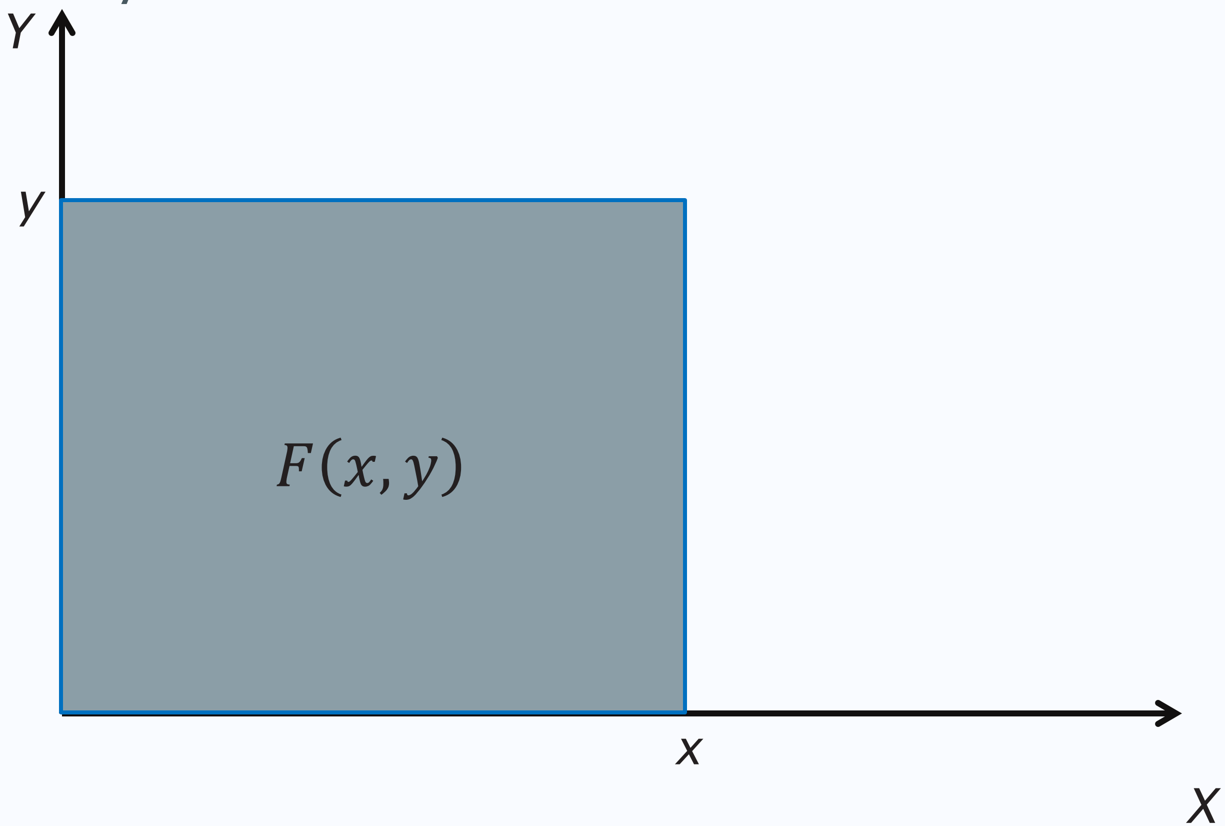
Multivariate Models

- Multivariate models extend this definition to cases where multiple quantities are observed at the same time
- The most simplest case is where two quantities X and Y are observed
- Now we have a two dimensional probability distribution function and a corresponding cumulative distribution

$$F(x, y) = P(X \leq x, Y \leq y)$$

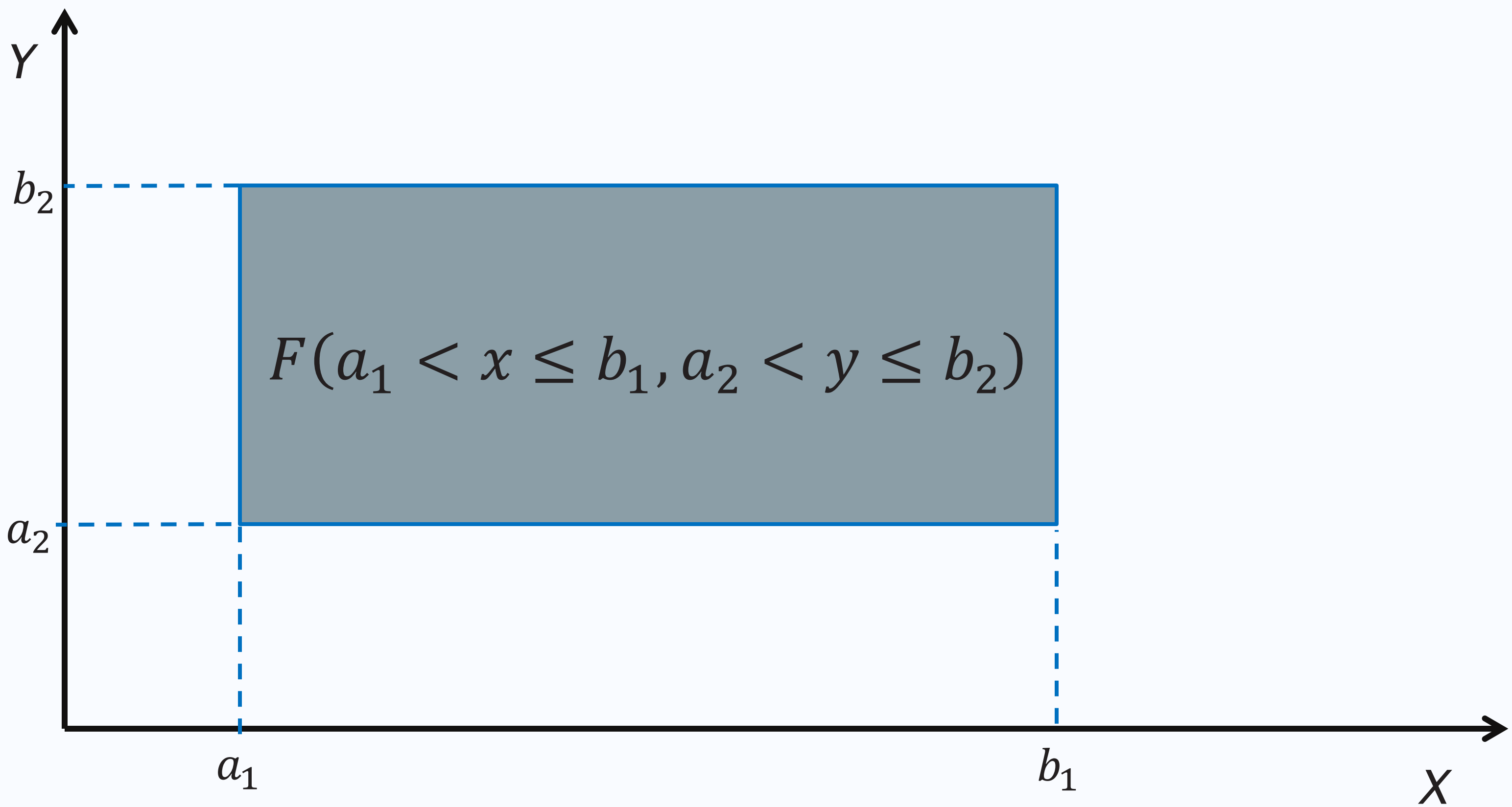
Multivariate Models

- Graphically this means:



Multivariate Models

- How do we extend our calculation between any set of bounds to multiple dimensions?

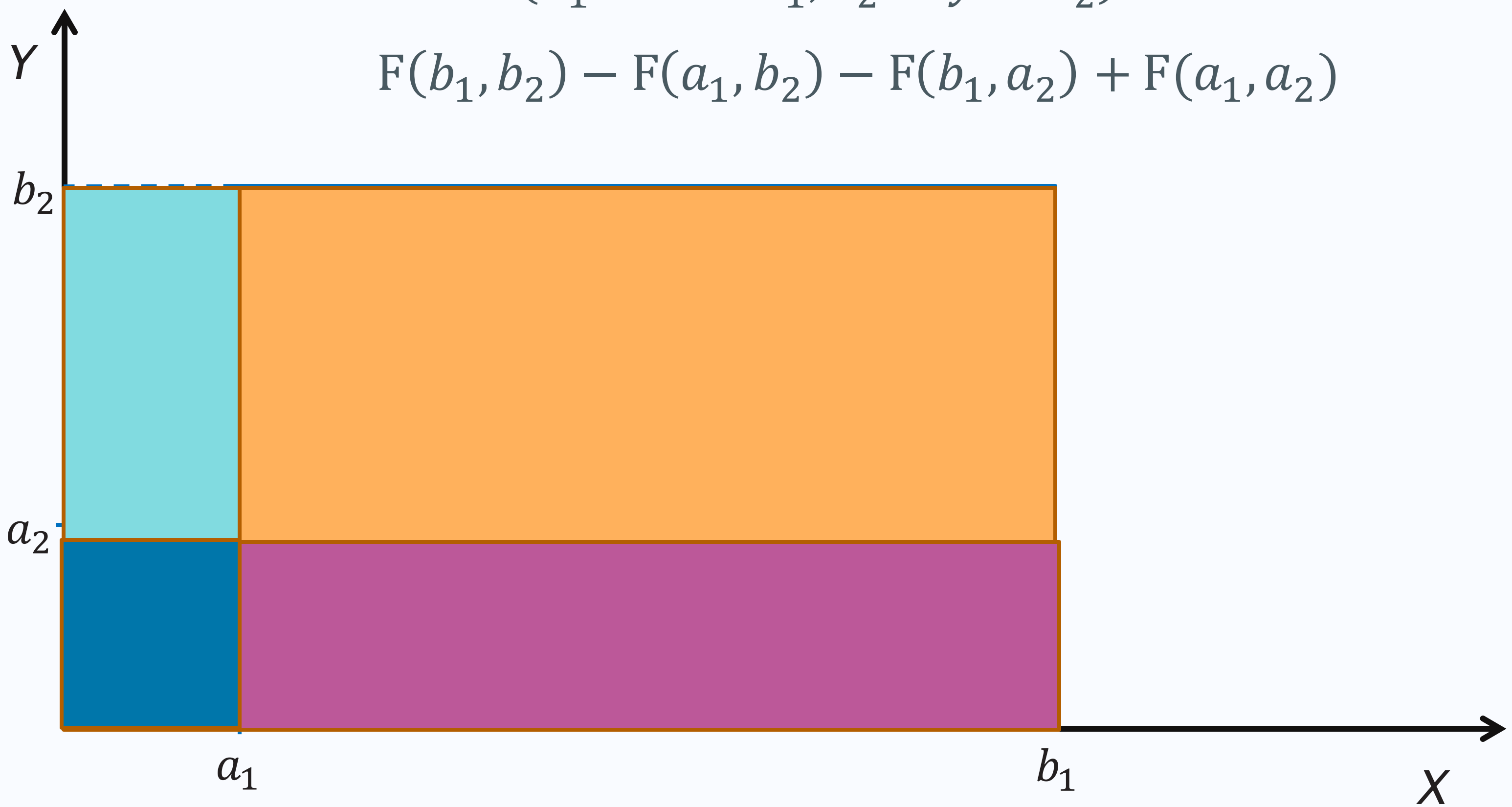


Multivariate Models

- We need to subdivide the problem into a series of areas which we do know

$$F(a_1 < x \leq b_1, a_2 < y \leq b_2) = \cdots$$

$$\mathrm{F}(b_1, b_2) - \mathrm{F}(a_1, b_2) - \mathrm{F}(b_1, a_2) + \mathrm{F}(a_1, a_2)$$



Multivariate Models

- Let's denote our univariate PDF as

$$f(x)$$

- Hence our two dimensional PDF is denoted as

$$f(x_1, x_2)$$

- With a corresponding CDF

$$F(a,b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x_1, x_2) dx_1 dx_2$$

- As per a univariate model

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_1, x_2) dx_1 dx_2 = 1$$

Multivariate Models

- Extending this definition to higher dimensions, our PDF is

$$f(x_1, x_2, \dots, x_n)$$

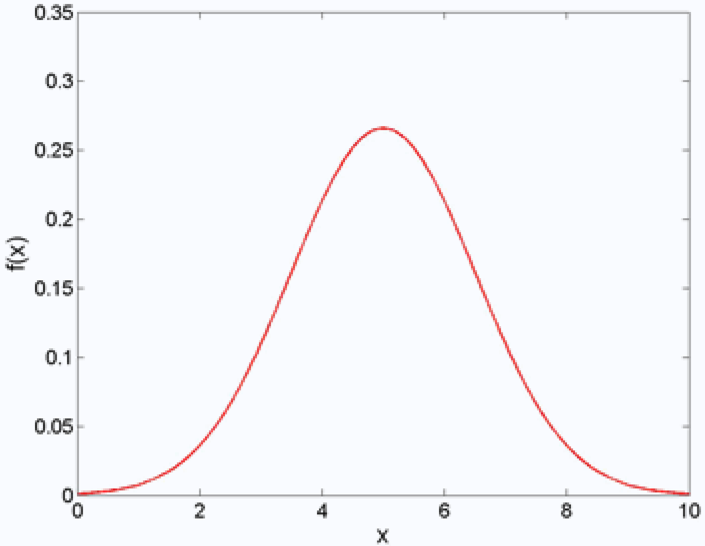
- Our CDF now involves a much more complicated integral

$$F(a_1, a_2, \dots, a_n) = \int_{-\infty}^{a_n} \dots \int_{-\infty}^{a_2} \int_{-\infty}^{a_1} f(x_1, x_2, \dots, x_n)\, dx_1 dx_2 \dots dx_n$$

- But this integral is always equal to 1 between $\pm\infty$

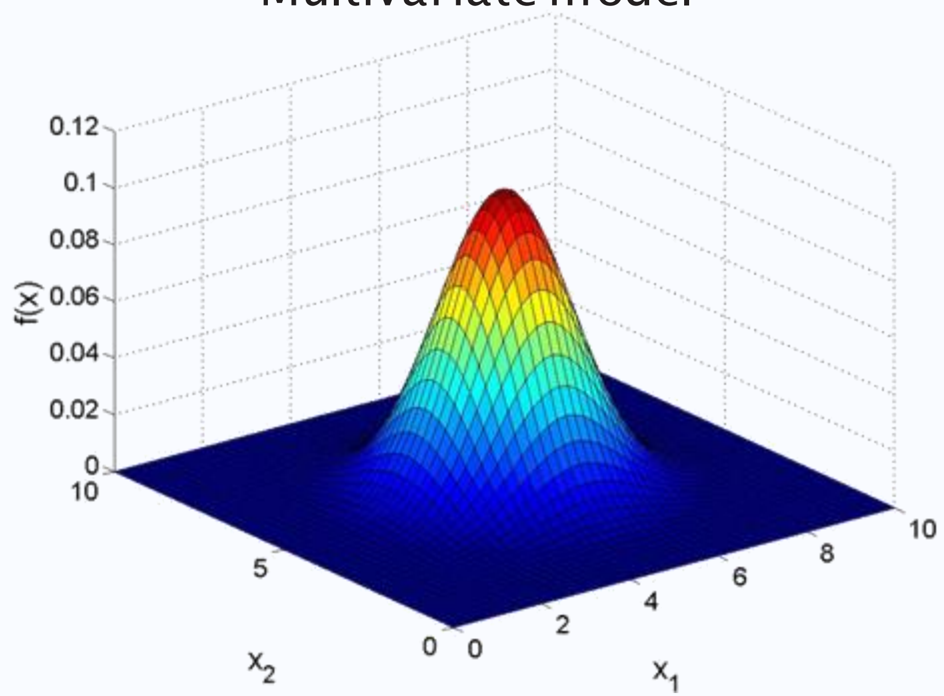
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, x_2, \dots, x_n)\, dx_1 dx_2 \dots dx_n = 1$$

Multivariate Normal Distribution



Univariate model

Multivariate model



Multivariate Normal Distribution

- Univariate PDF:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- Multivariate PDF:

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p|\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

Multivariate Normal Distribution

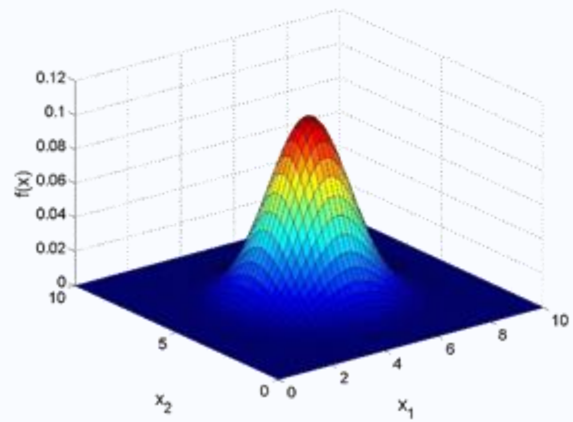
$$f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

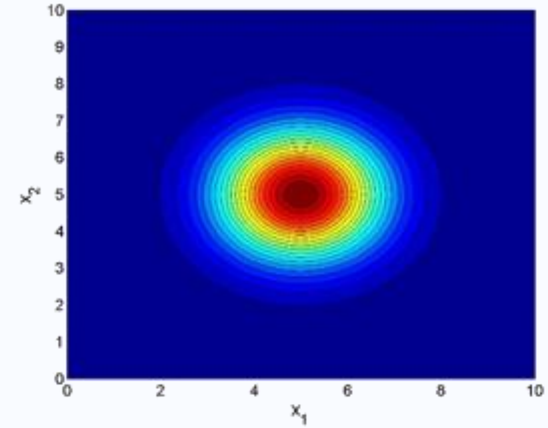
- $\boldsymbol{x}$ and $\boldsymbol{\mu}$ are now both vectors
- $\boldsymbol{\Sigma}$ defines the covariance matrix
- p defines the number of variables
- $|\boldsymbol{\Sigma}|$ defines the determinant of $\boldsymbol{\Sigma}$
- Clearly when $p = 1$ this equation reduces down to that of the univariate normal distribution

Multivariate Normal Distribution

$$\boldsymbol{\mu} = \begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$$

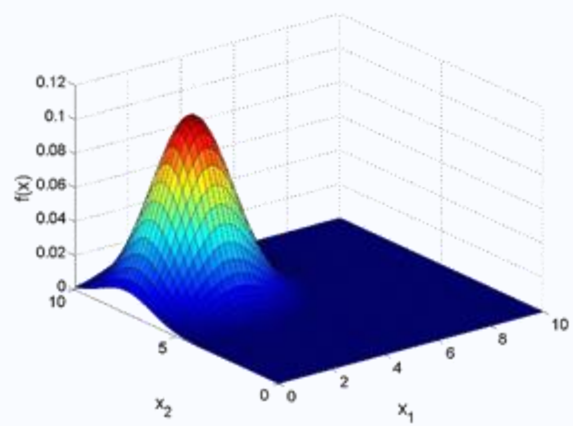
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.5 & 0.0 \\ 0.0 & 1.5 \end{bmatrix}$$

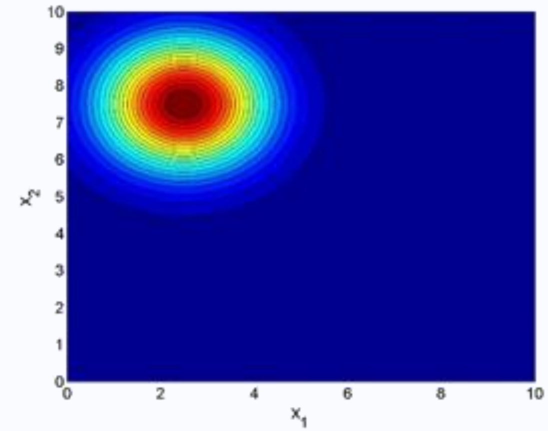




$$\boldsymbol{\mu} = \begin{bmatrix} 2.5 \\ 7.5 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.5 & 0.0 \\ 0.0 & 1.5 \end{bmatrix}$$

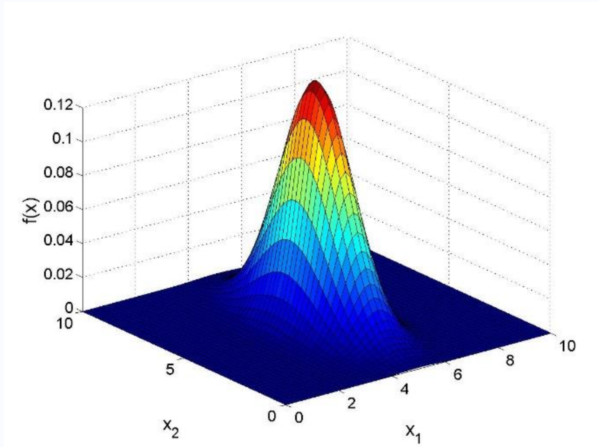


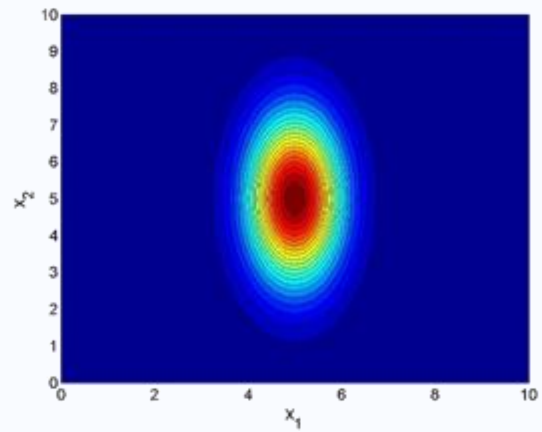


Multivariate Normal Distribution

$$\boldsymbol{\mu} = \begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$$

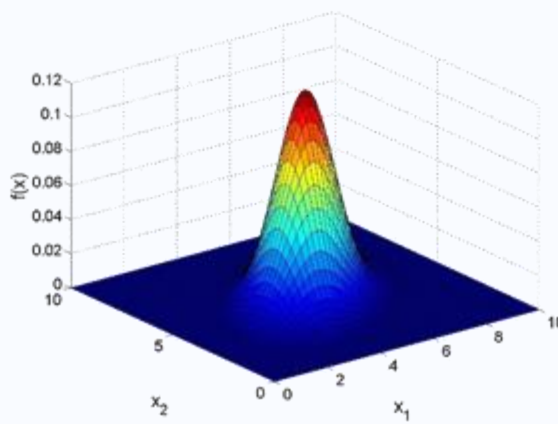
$$\boldsymbol{\Sigma} = \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 2.5 \end{bmatrix}$$

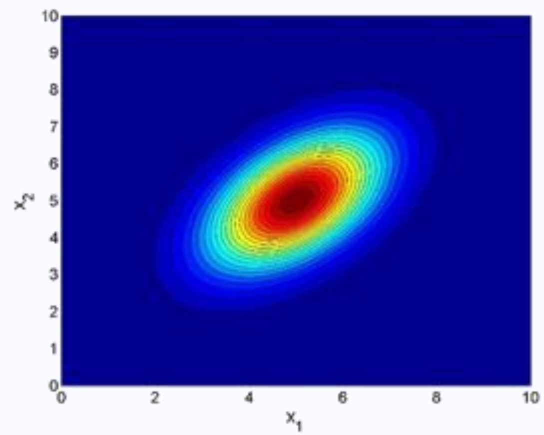




$$\boldsymbol{\mu} = \begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.5 & 0.75 \\ 0.75 & 1.5 \end{bmatrix}$$

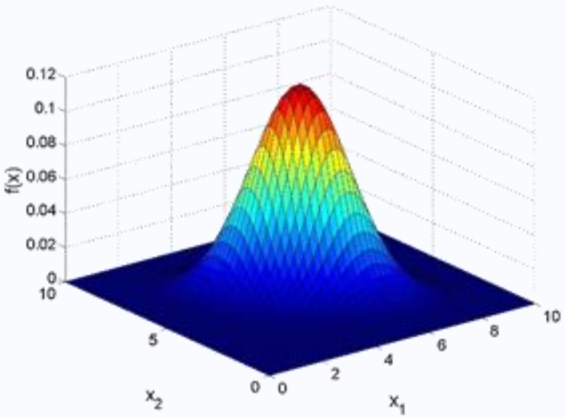


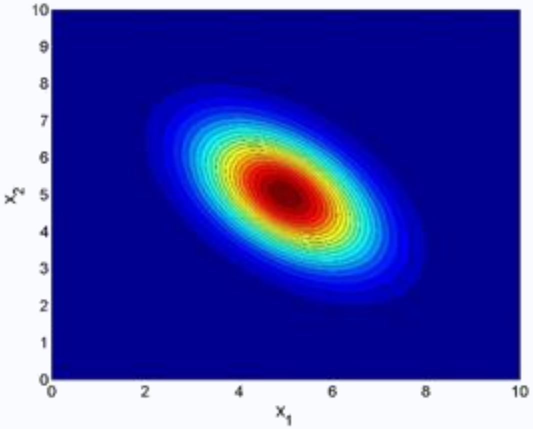


Multivariate Normal Distribution

$$\boldsymbol{\mu} = \begin{bmatrix} 5.0 \\ 5.0 \end{bmatrix}$$

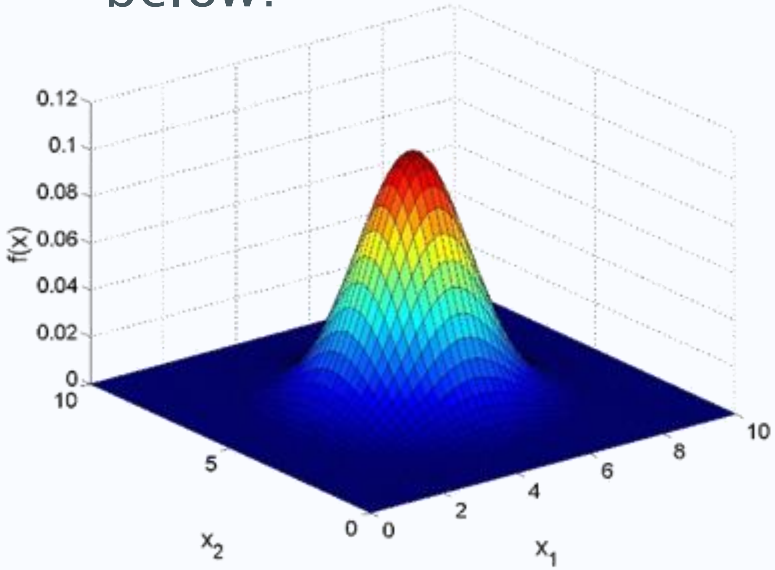
$$\boldsymbol{\Sigma} = \begin{bmatrix} 1.5 & -0.75 \\ -0.75 & 1.5 \end{bmatrix}$$



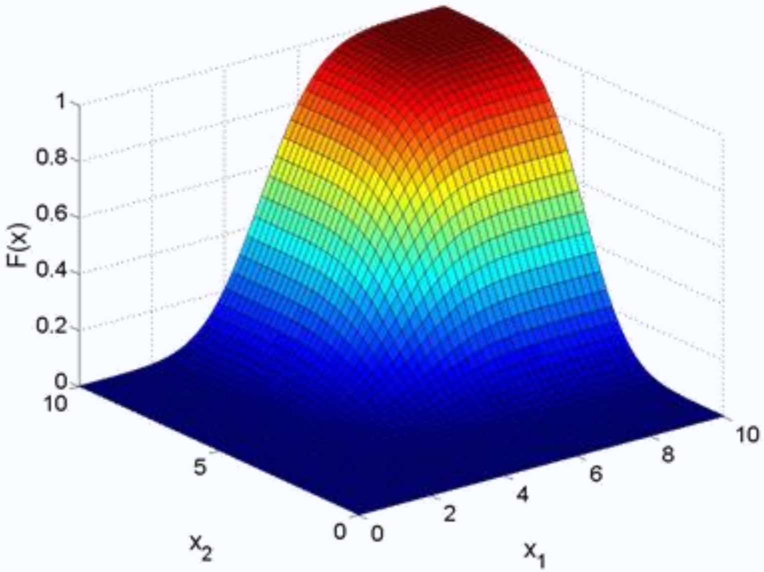


Multivariate Normal Distribution

- What about the CDF, what would it look like for the case below?



PDF



CDF

Multivariate Normal Distribution

- Matlab contains useful functions to calculate multivariate normal distribution PDFs and CDFs

```
mvnpdf(X,MU,SIGMA)

mvncdf(X,MU,SIGMA)
```

- In Python the scipy library contains a multivariate normal distribution function

```
scipy.stats.multivariate_normal
```

- There is no analytical expression for the CDF
  - Typically some form of quadrature is employed to calculate CDFs for multivariate models

Fitting a Multivariate Normal Distribution

- How can we fit such a distribution?
  - By maximum likelihood estimation of course!
- The PDF for a multivariate normal distribution depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
- $\boldsymbol{\mu}$ is a vector of length p giving us a first set of parameters to determine
- $\boldsymbol{\Sigma}$ is a symmetric matrix $p \times p$ in size giving us a further $\frac{1}{2}p(p+1)$ parameters
- In total this gives us $\frac{1}{2}p(p+3)$ parameters to define

Fitting a Multivariate Normal Distribution

- Clearly this is a lot of parameters as p increases
- Let's derive our likelihood function

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$

- Our expression for the PDF is

$$f(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^p |\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right]$$

Fitting a Multivariate Normal Distribution

- Which when combined gives us the following expression for the likelihood function

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = 2\pi^{\frac{-np}{2}} |\boldsymbol{\Sigma}|^{\frac{-1}{2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}_i - \boldsymbol{\mu})\right]$$

- Taking logs of this function we obtain

$$l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{-np}{2}\ln(2\pi) - \frac{n}{2}\ln(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

Fitting a Multivariate Normal Distribution

- To simplify things instead of maximising the likelihood let's minimise -2 times the likelihood

$$-2l(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = np\ln(2\pi) + n\ln(|\boldsymbol{\Sigma}|) + \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

- The first part of this expression is of course a constant independent of both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ which reduces our expression to

$$n\ln(|\boldsymbol{\Sigma}|) + \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-\mathbf{1}}(\boldsymbol{x}_i - \boldsymbol{\mu})$$

Fitting a Multivariate Normal Distribution

- Now we have a function to minimise we could use an optimisation algorithm to find the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$
- Fortunately there is a closed form solution for this problem
- The MLE $\boldsymbol{\mu}$ is equal to the sample mean for each set of observations

$$\widehat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}$$

$$\widehat{\boldsymbol{\mu}} = [\overline{\boldsymbol{x}}_1, \overline{\boldsymbol{x}}_2, \dots, \overline{\boldsymbol{x}}_n]$$

- Where:

$$\mu_i = \overline{\boldsymbol{x}}_i = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Fitting a Multivariate Normal Distribution

- The MLE of $\boldsymbol{\Sigma}$ is equal to the sample covariance matrix

$$\boldsymbol{\Sigma}_{\boldsymbol{ij}} = \mathrm{cov}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

- Where

$$\mathrm{cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (\boldsymbol{x}_i - \mu_i)^T (\boldsymbol{x}_j - \mu_j)/n$$

- The derivation of this closed form solution is beyond the scope of this module but it is in the literature for those interested

Joint Distribution Functions

- Assuming independence of the parameters it is possible to derive a multivariate joint distribution function of any form
- This is subject to a few restrictions:

  $F(-\infty, -\infty) = 0$

  $F(\infty, \infty) = 1$

  if a<b and c<d then $F(a, c) < F(b, d)$

- Of course, $\pm\infty$ bounds depend on the constituent distributions
  - This may not be appropriate for all cases e.g. an exponential

A Bivariate Exponential Distribution

- Let's define a bivariate exponential distribution function
- Recall that for the univariate exponential function

$$F(x) = 1 - \exp(-\lambda x)$$

- Our bivariate function (assuming statistical independence) is of the form

$$F(x_1, x_2) = F_1(x_1)F_2(x_2)$$

$$\therefore F(x_1, x_2) = \begin{cases} [1 - \exp(-\lambda_1 x_1)][1 - \exp(-\lambda_2 x_2)] & x_{1,2} \geq 0 \\ 0 & x_{1,2} < 0 \end{cases}$$

A Bivariate Exponential Distribution

- The PDF is therefore?

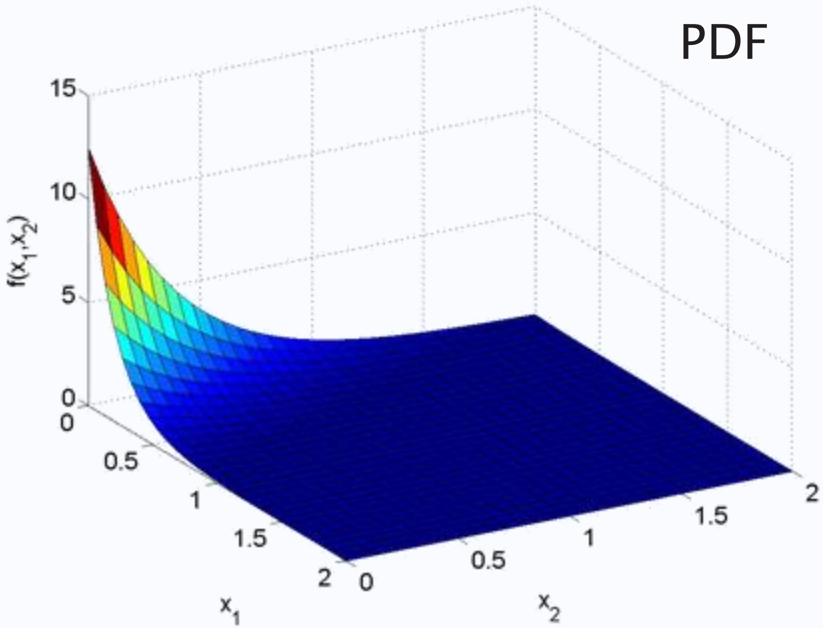
$$f(x_1, x_2) = \frac{\partial^2 F}{\partial x_1 \partial x_2}$$

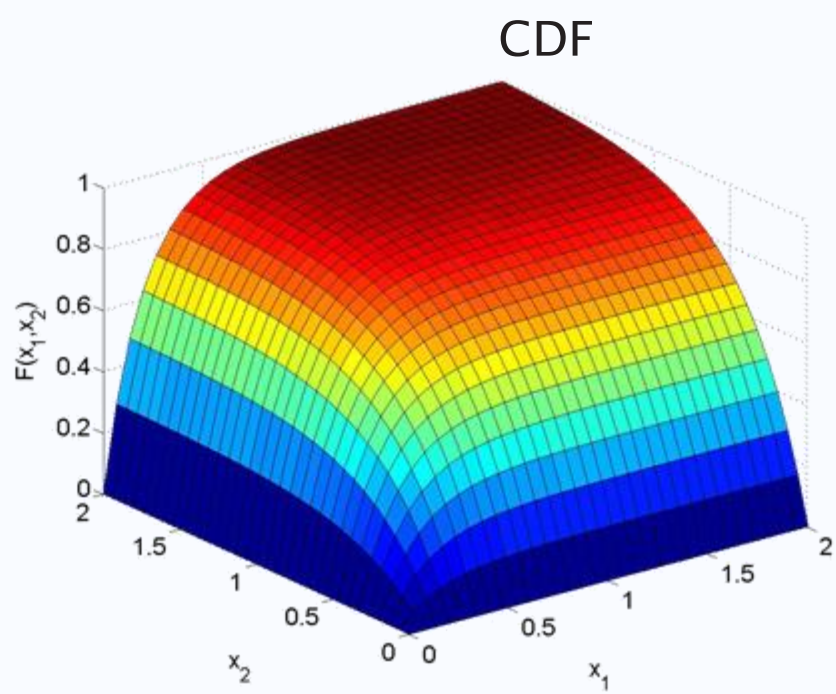
$$f(x_1, x_2) = \begin{cases} \lambda_1 \lambda_2 \exp(-\lambda_1 x_1 - \lambda_2 x_2) & x_{1,2} \geq 0 \\ 0 & x_{1,2} < 0 \end{cases}$$

- And the parameters λ_1 and λ_2 are found by MLE as normal

A Bivariate Exponential Distribution

- Assuming $\lambda_1 = 5$ and $\lambda_2 = 2.5$





Joint Distribution Functions

- Of course, there is nothing to stop you combining multiple different distribution functions in this manner e.g.
  - An exponential and a normal
  - Normal and log-normal
  - Or any combination
- For cases with >2 variables

$$F(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_i(x_i)$$

$$f(x_1, x_2, \ldots, x_n) = \frac{\partial^n F}{\partial x_1 \partial x_2 \ldots \partial x_n}$$

Joint Distribution Functions

- Of course, even in *n* dimensions….
  - We can still apply MLE to determine the optimal parameters for our model

$$L(\theta) = \prod_{i=1}^{n} f(t_i; \theta)$$

  - We can still use the Fisher information matrix to calculate confidence bounds in our parameters

$$I_{ij} = E\left[-\frac{\partial^2 l(t;\theta)}{\partial\theta_i \partial\theta_j}\right]$$

University of Southampton