• Probability density function (pdf) for a continuous random variable $X$

$$P(a \leq X \leq b) = \int_{a}^{b} p(x)dx$$

therefore

$$P(x \leq X \leq x + \delta x) \approx p(x)\delta x$$

• **Example:** Gaussian distribution

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

shorthand notation $X \sim N(\mu, \sigma^2)$

• Standard normal (or Gaussian) distribution $Z \sim N(0, 1)$

• Normalization

$$\int_{-\infty}^{\infty} p(x)dx = 1$$
The Gaussian density function is given by:

\[ p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{- (x - \mu)^2}{2\sigma^2}\right) \]
GAUSSIAN ≠ MULTIMODAL

\[ p(a \leq x \leq b) = \int_{a}^{b} p(u) \, du \]
GAUSSIAN IS A PARAMETRIC DISTRIBUTION

- Expectation
  $$E[g(X)] = \int g(x)p(x)dx$$
- mean, $E[X]$
- Variance $E[(X - \mu)^2]$
- For a Gaussian, mean $= \mu$, variance $= \sigma^2$
- Shorthand: $x \sim N(\mu, \sigma^2)$
PARAMETRIC DESCRIPTION HAS PROBLEMS WITH MULTIMODAL
BIVARIATE GAUSSIAN

• Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$

• If $X_1$ and $X_2$ are independent

$$p(x_1, x_2) = \frac{1}{2\pi(\sigma_1^2 \sigma_2^2)^{1/2}} \exp\left(-\frac{1}{2} \left\{ \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} \right\} \right)$$

• Let $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$, $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$

$$p(x) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left(-\frac{1}{2} \left\{ (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \right)$$
BIVARIATE JOINT PROBABILITY AND SAMPLING

\[ P(a \leq x \leq b, \ c \leq y \leq d) = \int_{c}^{d} \int_{a}^{b} p(x,y) \, dx \, dy \]
Bivariate is a joint density, that includes conditional

\[ P(a \leq x \leq b, c \leq y \leq d) = \int_c^d \int_a^b p(x, y) \, dx \, dy \]

\[ P(a \leq x \leq b \mid y) = \frac{\int_a^b p(u, y) \, du}{\int_{-\infty}^{+\infty} p(u, y) \, du} \]
Right: Distribution - Left: Contour curve of the distribution: an ellipsoid in the $x - y$ plane
• Covariance
• $\Sigma$ is the covariance matrix
\[ \Sigma = E[(x - \mu)(x - \mu)^T] \]
\[ \Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)] \]
• Example: plot of weight vs height for a population
BIVARIATE EXAMPLES OF DIFFERENT COVARIANCES

- $\mu = [0; 0]$
- $\Sigma = [1 \ 0; 0 \ 1]$

- $\mu = [0; 0]$
- $\Sigma = [1 \ 0.5; 0.5 \ 1]$

- $\mu = [0; 0]$
- $\Sigma = [1 \ 0.8; 0.8 \ 1]$
MULTIVARIATE GAUSSIAN

- $P(x \in \mathcal{R}) = \int_{\mathcal{R}} p(x) dx$

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

- $\Sigma$ is the covariance matrix
  
  $\Sigma = E[(x - \mu)(x - \mu)^T]$
  
  $\Sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$
MULTIVARIATE COVARIANCE

- $\Sigma$ is symmetric

- Shorthand $x \sim N(\mu, \Sigma)$

- For $p(x)$ to be a density, $\Sigma$ must be positive definite

- $\Sigma$ has $d(d + 1)/2$ parameters, the mean has a further $d$
THE CASE/EFFECT OF A DIAGONAL COVARIANCE MATRIX

To get an intuition for what a multivariate Gaussian is, consider the simple case where $n = 2$, and where the covariance matrix $\Sigma$ is diagonal, i.e.,

$$
\begin{bmatrix}
x_1 \\
x_2
\end{bmatrix}
\quad \quad 
\begin{bmatrix}
\mu_1 \\
\mu_2
\end{bmatrix}
\quad \quad 
\begin{bmatrix}
\sigma_1^2 & 0 \\
0 & \sigma_2^2
\end{bmatrix}
$$

In this case, the multivariate Gaussian density has the form,

$$
p(x; \mu, \Sigma) = \frac{1}{2\pi \begin{vmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{vmatrix}^{1/2}} \exp\left( -\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}^{-1} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right)
$$

$$
= \frac{1}{2\pi(\sigma_1^2 \cdot \sigma_2^2 - 0 \cdot 0)^{1/2}} \exp\left( -\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix} \right),
$$

where we have relied on the explicit formula for the determinant of a $2 \times 2$ matrix$^3$, and the fact that the inverse of a diagonal matrix is simply found by taking the reciprocal of each diagonal entry. Continuing,

$$
p(x; \mu, \Sigma) = \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left( -\frac{1}{2} \begin{bmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{bmatrix}^T \begin{bmatrix} \frac{1}{\sigma_1^2} (x_1 - \mu_1) \\ \frac{1}{\sigma_2^2} (x_2 - \mu_2) \end{bmatrix} \right)
$$

$$
= \frac{1}{2\pi \sigma_1 \sigma_2} \exp\left( -\frac{1}{2\sigma_1^2} (x_1 - \mu_1)^2 - \frac{1}{2\sigma_2^2} (x_2 - \mu_2)^2 \right)
$$

$$
= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left( -\frac{1}{2\sigma_1^2} (x_1 - \mu_1)^2 \right) \cdot \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left( -\frac{1}{2\sigma_2^2} (x_2 - \mu_2)^2 \right).
$$

The last equation we recognize to simply be the product of two independent Gaussian densities, one with mean $\mu_1$ and variance $\sigma_1^2$, and the other with mean $\mu_2$ and variance $\sigma_2^2$.

More generally, one can show that an $n$-dimensional Gaussian with mean $\mu \in \mathbb{R}^n$ and diagonal covariance matrix $\Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$ is the same as a collection of $n$ independent Gaussian random variables with mean $\mu_i$ and variance $\sigma_i^2$, respectively.
$d^2_{\Sigma}(x_i, x_j) = (x_i - x_j)^T \Sigma^{-1} (x_i - x_j)$

- $d^2_{\Sigma}(x_i, x_j)$ is called the Mahalanobis distance between $x_i$ and $x_j$
- If $\Sigma$ is diagonal, the contours of $d^2_{\Sigma}$ are axis-aligned ellipsoids
- If $\Sigma$ is not diagonal, the contours of $d^2_{\Sigma}$ are *rotated* ellipsoids

$$\Sigma = U \Lambda U^T$$

where $\Lambda$ is diagonal and $U$ is a rotation matrix

- $\Sigma$ is positive definite $\Rightarrow$ entries in $\Lambda$ are positive
Euclidian distance weights all dimensions (variables) equally, however, statistically they may not be the same:

The Euclidian distance tells us that \((x_2 - x_1) = (y_2 - y_1)\), however statistically \((x_2 - x_1) < (y_2 - y_1)\).
Mahalanobis Distance

It is easy to see that for low (or zero) covariance we can normalise the distances by dividing by the variance:

\[ \Delta'x = (x_2 - x_1) / \sqrt{\sigma_{xx}} \]
\[ \Delta'y = (y_2 - y_1) / \sqrt{\sigma_{yy}} \]

Mahalanobis Distance = \( \sqrt{(\Delta'x^2 + \Delta'y^2)} \)

In general the Mahalanobis distance between two points can be written as:

\[ \sqrt{((x_2 - x_1, y_2 - y_1) \Sigma^{-1} (x_2 - x_1, y_2 - y_1))} \]
Mahalanobis Distance

The Mahalanobis distance also works for high co-variance.

For the 2D case the inverse of the covariance matrix is:

$$\Sigma^{-1} = \frac{1}{(\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2)} \begin{pmatrix} \sigma_{yy} & -\sigma_{xy} \\ -\sigma_{xy} & \sigma_{xx} \end{pmatrix}$$
PARAMETRIZATION OF THE COVARIANCE MATRIX

- Fully general $\Sigma \implies$ variables are correlated

- Spherical or isotropic. $\Sigma = \sigma^2 I$. Variables are independent

- Diagonal $[\Sigma]_{ij} = \delta_{ij} \sigma_i^2$ Variables are independent

- Rank-constrained: $\Sigma = WW^T + \Psi$, with $W$ being a $d \times q$ matrix with $q < d - 1$ and $\Psi$ diagonal. This is the factor analysis model. If $\Psi = \sigma^2 I$, then with have the probabilistic principal components analysis (PPCA) model
• Linear transformations of Gaussian RVs are Gaussian
  \[ X \sim N(\mu_x, \Sigma) \]
  \[ Y = AX + b \]
  \[ Y \sim N(A\mu_x + b, A\Sigma A^T) \]

• Sums of Gaussian RVs are Gaussian
  \[ Z = X + Y \]
  \[ E[Z] = E[X] + E[Y] \]
  \[ \text{var}[Z] = \text{var}[X] + \text{var}[Y] + 2\text{covar}[XY] \]
  if \( X \) and \( Y \) are independent \( \text{var}[Z] = \text{var}[X] + \text{var}[Y] \)
• Gaussian has relatively simple analytical properties

• Central limit theorem. Sum (or mean) of $M$ independent random variables is distributed normally as $M \to \infty$ (subject to a few general conditions)

• Diagonalization of covariance matrix $\implies$ rotated variables are independent

• All marginal and conditional densities of a Gaussian are Gaussian

• The Gaussian is the distribution that maximizes the entropy $H = -\int p(x) \log p(x) dx$ for fixed mean and covariance
Consider a random vector \( x \in \mathbb{R}^n \) with \( x \sim \mathcal{N}(\mu, \Sigma) \). Suppose also that the variables in \( x \) have been partitioned into two sets \( x_A = [x_1 \, \cdots \, x_r]^T \in \mathbb{R}^r \) and \( x_B = [x_{r+1} \, \cdots \, x_n]^T \in \mathbb{R}^{n-r} \) (and similarly for \( \mu \) and \( \Sigma \)), such that

\[
x = \begin{bmatrix} x_A \\ x_B \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}.
\]

Here, \( \Sigma_{AB} = \Sigma_{BA}^T \) since \( \Sigma = E[(x - \mu)(x - \mu)^T] = \Sigma^T \). The following properties hold:

1. **Normalization.** The density function normalizes, i.e.,

\[
\int_x p(x; \mu, \Sigma)dx = 1.
\]

This property, though seemingly trivial at first glance, turns out to be immensely useful for evaluating all sorts of integrals, even ones which appear to have no relation to probability distributions at all (see Appendix A.1)!

2. **Marginalization.** The marginal densities,

\[
p(x_A) = \int_{x_B} p(x_A, x_B; \mu, \Sigma)dx_B
\]

\[
p(x_B) = \int_{x_A} p(x_A, x_B; \mu, \Sigma)dx_A
\]

are Gaussian:

\[
x_A \sim \mathcal{N}(\mu_A, \Sigma_{AA})
\]

\[
x_B \sim \mathcal{N}(\mu_B, \Sigma_{BB}).
\]
3. **Conditioning.** The conditional densities

\[
p(x_A | x_B) = \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_A} p(x_A, x_B; \mu, \Sigma) dx_A}
\]

\[
p(x_B | x_A) = \frac{p(x_A, x_B; \mu, \Sigma)}{\int_{x_B} p(x_A, x_B; \mu, \Sigma) dx_B}
\]

are also Gaussian:

\[
x_A | x_B \sim \mathcal{N}\left(\mu_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \mu_B), \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}\right)
\]

\[
x_B | x_A \sim \mathcal{N}\left(\mu_B + \Sigma_{BA}\Sigma_{AA}^{-1}(x_A - \mu_A), \Sigma_{BB} - \Sigma_{BA}\Sigma_{AA}^{-1}\Sigma_{AB}\right).
\]

A proof of this property is given in Appendix A.2.

4. **Summation.** The sum of independent Gaussian random variables (with the same dimensionality), \(y \sim \mathcal{N}(\mu, \Sigma)\) and \(z \sim \mathcal{N}(\mu', \Sigma')\), is also Gaussian:

\[
y + z \sim \mathcal{N}(\mu + \mu', \Sigma + \Sigma').
\]