Lecture 3

Review the fundamentals of probability

- Definition of probability
- Rules of probability
- Bayes' rule

Pdfs of real-valued quantities

Pdf characterisations

- Expectations, Covariance matrices

Gaussian distributions

When are probabilities used

Probabilities are used to describe two quantities

- *Frequencies of outcomes in random experiments.*
  - The probability of a coin toss landing as tails is $\frac{1}{2}$. Then if the coin is tossed n times and k_n "tails" are observed, it is expected $\frac{k_n}{n} \to \frac{1}{2}$ as $n \to \infty$.
- *Degree of belief in propositions not involving random variables.*
  - the probability that Mr S. was the murderer of Mrs S. given the evidence
  - the probability that this image contains a car given a calculated feature vector.

Defining probability

Define a probabilistic ensemble with a triple $(x, \mathcal{A}_X, \mathcal{P}_X)$, where x is the outcome of a random variable, X, and takes on one of a set of possible values, $\mathcal{A}_X = (a_1, a_2, \ldots, a_I)$, having probabilities $\mathcal{P}_X = (p_1, p_2, \ldots, p_I)$ with $P(X = a_i) = p_i$.

The following must be satisfied:

- $p_i \geq 0$ for $i = 1, \ldots, I$
- $\sum_{x \in \mathcal{A}_X} P(X = x) = 1$.

A simple example

Let x be the outcome of **throwing an unbiased die**, then

$$\mathcal{A}_X = \{`1`, `2`, `3`, `4`, `5`, `6`\}$$

$$\mathcal{P}_X = \left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}$$

Question:

$$P(x = `3`) = ?$$

$$P(x = `5`) = ?$$

Definitions of probability

Probability of a subset: If $V \subset \mathcal{A}_X$, then

$$P(V) = P(x \in V) = \sum_{x \in V} P(x)$$

Example:

Going back to our die example, let $V = \{`2`, `3`, `4`\}$, then

$$\begin{aligned} P(V) &= P(x = `2`) + P(x = `3`) + P(x = `4`) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2} \end{aligned}$$

The simple example

Throwing an unbiased die

$$\mathcal{A}_X = \{\text{`}1\text{`}, \text{`}2\text{`}, \text{`}3\text{`}, \text{`}4\text{`}, \text{`}5\text{`}, \text{`}6\text{`}\}$$

$$\mathcal{P}_X = \left\{\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}\right\}$$

Question:

If $V = \{\text{`}2\text{`}, \text{`}3\text{`}\}$, what is $P(V)$?

Definitions of probability

Joint probability: $X \times Y$ is an ensemble in which an outcome is an ordered pair (x, y) with $x \in \mathcal{A}_X = \{a_1, \ldots, a_I\}$ and $y \in \mathcal{B}_Y = \{b_1, \ldots, b_J\}$. Then $P(x, y)$ is the joint probability of x and y.

Example:

Remember the outcome of throwing an unbiased die is described with

$$\mathcal{A}_X = \underbrace{\{‘1‘, ‘2‘, ‘3‘, ‘4‘, ‘5‘, ‘6‘\}}_{\text{Possible outcomes}}, \qquad \mathcal{P}_X = \underbrace{\{6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}\}}_{\text{Probability of each outcome}}$$

Definitions of probability

Example ctd:

The output of **two consecutive** independent, T_1 and T_2 throws of an unbiased die:

Throw 1:

$$\begin{aligned} \mathcal{A}_{T_1} &= \{`1`, `2`, `3`, `4`, `5`, `6`\} \\ \mathcal{P}_{T_1} &= \{6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}\} \end{aligned}$$

Throw 2:

$$\begin{aligned} \mathcal{A}_{T_2} &= \{`1`, `2`, `3`, `4`, `5`, `6`\} \\ \mathcal{P}_{T_2} &= \{6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}, 6^{-1}\} \end{aligned}$$

Possible outcomes:

$$\begin{aligned} \mathcal{A}_{T_1 \times T_2} = \{ & (`1`, `1`), (`1`, `2`), (`1`, `3`), (`1`, `4`), (`1`, `5`), (`1`, `6`), \\ & (`2`, `1`), (`2`, `2`), (`2`, `3`), (`2`, `4`), (`2`, `5`), (`2`, `6`), \\ & \qquad \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ & (`6`, `1`), (`6`, `2`), (`6`, `3`), (`6`, `4`), (`6`, `5`), (`6`, `6`)\} \end{aligned}$$

Probabilities:

$$\mathcal{P}_{T_1 \times T_2} = \left\{\tfrac{1}{36}, \tfrac{1}{36}, \cdots, \tfrac{1}{36}\right\}$$

Another example

Scenario:

A person throws an **unbiased** die. If the outcome is **even** throw this die again, otherwise throw a die biased towards ‘3‘ with

$$\mathcal{P}_X = \left\{ \frac{1}{10}, \frac{1}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10} \right\}$$

Questions:

What is the set, $\mathcal{A}_{T_1 \times T_2}$, of possible outcomes?
What are the values in $\mathcal{P}_{T_1 \times T_2}$?

Definitions of probability

Marginal probability:

$$P(x = a_i) \equiv \sum_{y \in \mathcal{A}_Y} P(x = a_i, y)$$

Similarly:

$$P(y = b_j) \equiv \sum_{x \in \mathcal{A}_X} P(x, y = b_j)$$

Example:

Returning to example modelling the output of two consecutive independent throws of an unbiased die then

$$P(t_1 = `1`) = \sum_{i=1}^{6} P(t_1 = `1`, t_2 = `i`) = \sum_{i=1}^{6} \frac{1}{36} = \frac{1}{6}$$

Example

Scenario:

A person throws an unbiased die. If the outcome is even, throw this die again, otherwise throw a die biased towards '3' with

$$\mathcal{P}_X = \left\{ \frac{1}{10}, \frac{1}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10} \right\}$$

Question:

Given $P(t_1, t_2)$ (ie $\mathcal{P}_{T_1 \times T_2}$) and the defintion of marginal probability, calculate $P(t_2)$ the probability of the output of the second die in this scenario.

Definitions of probability

Conditional probability:

$$P(X = a_i \,|\, Y = b_j) = \frac{P(X = a_i, Y = b_j)}{P(Y = b_j)}, \text{ if } P(Y = b_j) \neq 0$$

Example:

Returning to example modelling the output of two consecutive independent throws of an unbiased die then

$$P(t_2 = \text{‘}3\text{‘} \,|\, t_1 = \text{‘}1\text{‘}) = \frac{P(t_1 = \text{‘}1\text{‘}, t_2 = \text{‘}3\text{‘})}{P(t_1 = \text{‘}1\text{‘})} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$$

Example

Scenario:

A person throws an unbiased die. If the outcome is even, throw this die again, otherwise throw a die biased towards $`3`$ with

$$\mathcal{P}_X = \left\{ \frac{1}{10}, \frac{1}{10}, \frac{1}{2}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10} \right\}$$

Question:

Calculate $P(t_2 = `3` \,|\, t_1 = `1`)$ and $P(t_2 = `3` \,|\, t_1 = `2`)$.

Rules of probability

Product Rule: from the definition of the conditional probability

$$P(x, y) = P(x \,|\, y)P(y) = P(y \,|\, x)P(x)$$

Sum/Chain Rule: rewriting the marginal probability definition

$$P(x) = \sum_y P(x, y) = \sum_y P(x \,|\, y)P(y)$$

Bayes' Rule: from the product rule

$$P(y|x) = \frac{P(x \,|\, y)P(y)}{P(x)} = \frac{P(x \,|\, y)P(y)}{\sum_{y'} P(x \,|\, y')P(y')}$$

Independence

Independence: Two random variables X, Y are independent if

$$P(x, y) = P(x)\, P(y) \quad \forall x \in \mathcal{A}_X, \forall y \in \mathcal{B}_Y.$$

This implies that

$$P(x \,|\, y) = P(x) \quad \forall x \in \mathcal{A}_X, \forall y \in \mathcal{B}_Y$$

and

$$P(y \,|\, x) = P(y) \quad \forall x \in \mathcal{A}_X, \forall y \in \mathcal{B}_Y.$$

X and Y are independent is often denoted by $X \perp\!\!\!\perp Y$.

An Example

Problem: Jo has the test for a nasty disease. Let a denote the state of Jo's health and b the test results.

$$a = \begin{cases} 1 & \text{if Jo has the disease,} \\ 0 & \text{Jo does not have the disease} \end{cases} \qquad b = \begin{cases} 1 & \text{if the test is positive,} \\ 0 & \text{if the test is negative.} \end{cases}$$

The test is 95% reliable, that is

$$\begin{aligned} p(b = 1 \,|\, a = 1) &= .95 \qquad & p(b = 1 \,|\, a = 0) &= .05 \\ p(b = 0 \,|\, a = 1) &= .05 \qquad & p(b = 0 \,|\, a = 0) &= .95 \end{aligned}$$

The final piece of background information is that 1% of people Jo's age and background have the disease.
Jo has the test and the result is **positive**.

What is the probability Jo **has** the disease?

Solution: The background information tells us

$$P(a = 1) = .01, \qquad P(a = 0) = .99$$

Jo would like to know how plausible it is that she has the disease. This involves calculating $P(a = 1 \,|\, b = 1)$ which is the probability of Jo having the disease given a positive test result.

Applying Bayes' Rule:

$$\begin{aligned}
P(a=1\,|\,b=1) &= \frac{P(b=1\,|\,a=1)P(a=1)}{P(b=1)} \\
&= \frac{P(b=1\,|\,a=1)P(a=1)}{P(b=1\,|\,a=1)P(a=1)+P(b=1\,|\,a=0)P(a=0)} \\
&= \frac{.95\times .01}{.95\times .01+.05\times .99} = .16
\end{aligned}$$

Your turn

Scenario: Your friend has two envelopes. One he calls the Win envelope which has 100 dollars and four beads (2 red and 2 blue) in it. While the other the $Lose$ envelope has three beads (1 red and 2 blue) and no money. You choose one of the envelopes at random and then your friend offers to sell it to you.

Question:

- How much should you pay for the envelope?
- Suppose before deciding you are allowed to draw one bead from the envelope.

  If this bead is blue how much should you pay?

Inference is important

Inference is the term given to the conclusions reached from the basis of evidence and reasoning.

Most of this course will be devoted to **inference** of some form.

Some examples:

I've got this evidence. What's the chance that this conclusion is true?

- I've got a sore neck: how likely am I to have meningitis
- My car detector has fired in this image: how likely is it there is a car in the image?

Inference using Bayes' rule

In general:

If $\boldsymbol{\theta}$ denotes the unknown parameters/decision, D the data and $\mathcal{H}$ denotes the overall hypothesis space, then

$$p(\boldsymbol{\theta} \,|\, D, \mathcal{H}) = \frac{P(D \,|\, \boldsymbol{\theta}, \mathcal{H}) P(\boldsymbol{\theta}|\mathcal{H})}{P(D \,|\, \mathcal{H})}$$

is written as

$$\text{posterior} = \frac{\text{likelihood} \;\times\; \text{prior}}{\text{evidence}}$$

Bayesian classification

Bayes' Rule can be expressed as

$$P(\omega_j \,|\, \mathbf{x}) = \frac{P(\mathbf{x} \,|\, \omega_j) P(\omega_j)}{\sum_{k=1}^{N} P(\mathbf{x} \,|\, \omega_k) P(\omega_k)} = \frac{P(\mathbf{x} \,|\, \omega_j) P(\omega_j)}{P(\mathbf{x})}$$

where ω_j is the jth class and $\mathbf{x}$ is the feature vector.

A typical decision rule (class assignment)

Choose the class ω_i with the highest $P(\omega_i \,|\, \mathbf{x})$. Intuitively, we will choose the class that is more *likely* given feature vector $\mathbf{x}$.

Terminology Each term in the Bayes' Rule has a special name:

$P(\omega_i)$ – **Prior probability** of class ω_i

$P(\omega_i \,|\, \mathbf{x})$ – **Posterior probability** of class ω_i given the observation $\mathbf{x}$

$P(\mathbf{x} \,|\, \omega_i)$ – **Likelihood** of observation $\mathbf{x}$ given class ω_i

$P(\mathbf{x})$ – **Evidence** the normalization constant

Bayes Classifier in a nutshell

1. Learn the class conditional distributions for each class ω_j.
2. This gives $P(\mathbf{x}\,|\,\omega_j)$
3. Estimate the prior $P(\omega)$ of each class
4. For a new data point $\mathbf{x}$ make a prediction with:

$$\omega^* = \arg\max_{\omega_j}\; P(\mathbf{x}\,|\,\omega_j)\,P(\omega_j)$$

Step one is know as **density estimation**. This will be the topic of several future lectures. We will also be examining the strengths and weaknesses of the Bayes classifiers.

We don't live in a purely discrete world

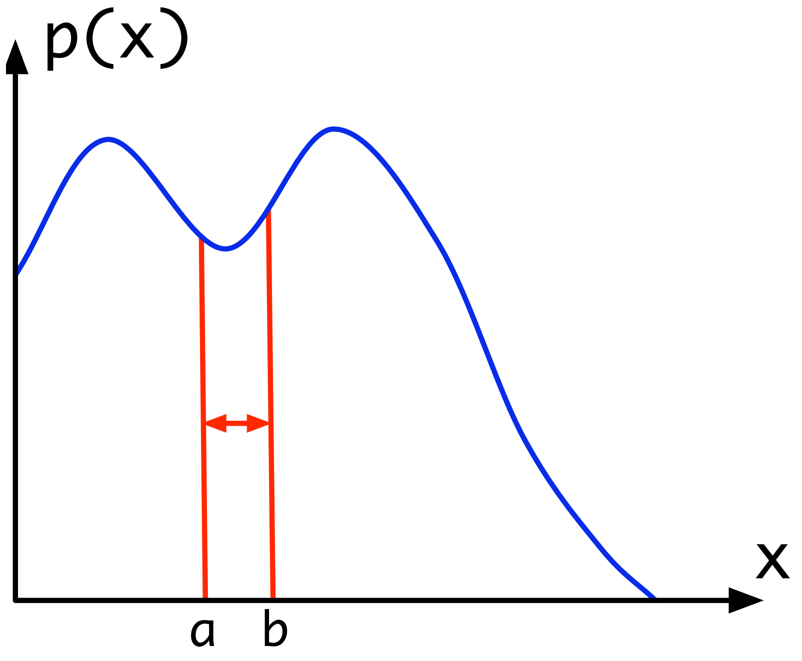
Continuous random variables

So far have only encountered discrete random variables. But the outcome x of the random variable can be continuous.

In this case $\mathcal{A}_X$ is an interval or union of intervals such as $\mathcal{A}_X = (-\infty, \infty)$. The notion of probability must also be updated. Now $p(\cdot)$ denotes the probability density function (pdf). It has the two properties:

$$\mathbf{1}) \quad p(x) \geq 0 \;\; \forall x \in \mathcal{A}_X,$$

$$\mathbf{2}) \quad \int_{x \in \mathcal{A}_X} p(x)\, dx = 1.$$

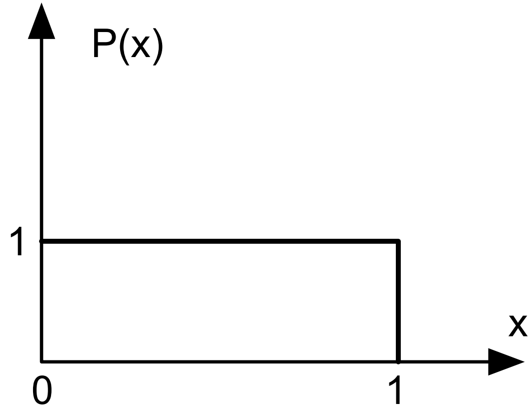


The probability that a continuous random variable x lies between values a and b (with $b > a$) is defined to be

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

Continuous random variables

An example of a continuous probability distribution function $p(\cdot)$:



$$p(x) = \begin{cases} 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

The above is known as the **uniform** density function.

Continuous random variables

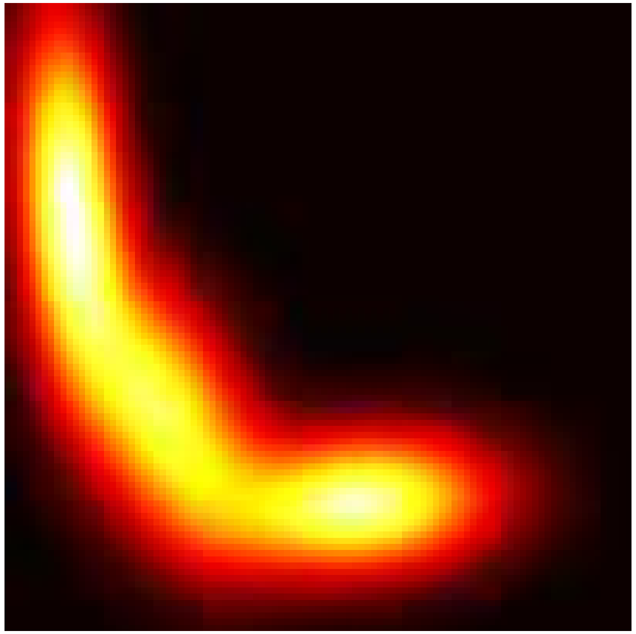
All the previous definitions and rules are the same except that the summation signs are replaced by integral signs where appropriate. For example:

Marginal probability

$$p(x) \equiv \int_{y \in \mathcal{A}_Y} p(x,y)\, dy$$

m dimensional random variables

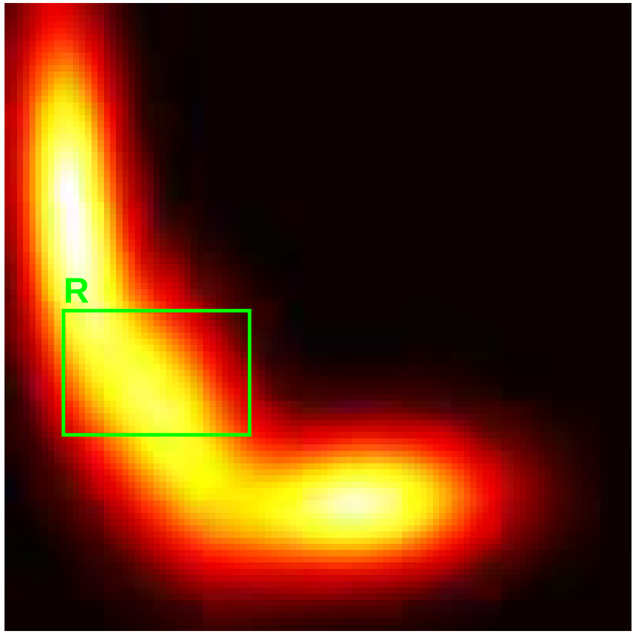
Below is he joint probability density of an ordered pair (X, Y) and $p(x, y)$.



Consider also the ordered vector $\mathbf{X} = (X_1, X_2, \ldots, X_m)$. The

Joint Probability Density for this vector is defined as $p(\mathbf{x})$ or $p(x_1, x_2, \ldots, x_m)$.

Probability of a volume



If R is a volume defined in the space of possible outcomes then the probability of this volume is

$$P((X_1, X_2, \ldots, X_m) \in R) = \underbrace{\int\int\cdots\int}_{(x_1,x_2,\ldots,x_m)\in R} p(x_1, x_2, \ldots, x_m)\, dx_1\, dx_2 \ldots dx_m$$

Marginal pdf

The marginal pdf is used to represent the pdf of a subset of the $x_i's$.

$$p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m) \equiv \int_z p(x_1, \ldots, x_{i-1}, x_i = z, x_{i+1}, \ldots, x_m)\, dz$$

PDF description

The probability density function, $p_X(\cdot)$, fully characterizes our random variable X. The following partially characterize $p_X(\cdot)$.

Expectation: $$\mathrm{E}\left[X\right] = \int_{x \in \mathcal{A}_X} x\, p(x)\, dx = \mu$$

represents the center of mass of the density.

Variance: $$\mathrm{Var}\left[X\right] = \mathrm{E}\left[(X - \mathrm{E}\left[X\right])^2\right] = \int_{x \in \mathcal{A}_X} (x - \mu)^2\, p(x)\, dx$$

represents the spread about the mean.

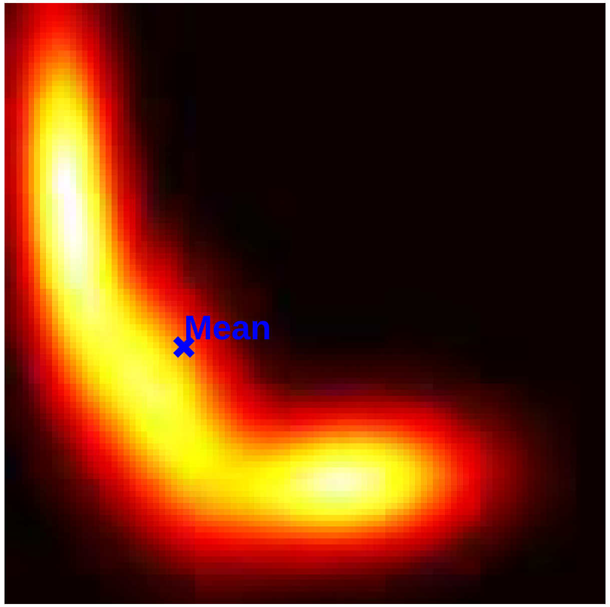
Std deviation: $\mathrm{std}\left[X\right] = \sqrt{\mathrm{Var}\left[X\right]}$, square root of the variance.

Note: $\mathrm{E}\left[f(X)\right] = \int_{x \in \mathcal{A}_X} f(x)\, p(x)\, dx$

Partial description

Mean vector

$$\mathrm{E}\left[\mathbf{X}\right] = \left(\mathrm{E}\left[X_1\right], \mathrm{E}\left[X_2\right], \ldots, \mathrm{E}\left[X_N\right]\right)^T = \left(\mu_1, \ldots, \mu_N\right)^T = \boldsymbol{\mu}$$



Covariance Matrix

$$\begin{aligned}
\mathrm{Cov}\left[\mathbf{X}\right] &= \mathrm{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right] \\
&= \begin{pmatrix} \mathrm{E}\left[(X_1-\mu_1)(X_1-\mu_1)\right] & \ldots & \mathrm{E}\left[(X_1-\mu_1)(X_N-\mu_N)\right] \\ \vdots & \cdots & \vdots \\ \mathrm{E}\left[(X_N-\mu_N)(X_1-\mu_1)\right] & \ldots & \mathrm{E}\left[(X_N-\mu_N)(X_N-\mu_N)\right] \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \cdots & c_{1N} \\ \cdots & \ddots & \cdots \\ c_{1N} & \cdots & \sigma_N^2 \end{pmatrix} \\
&= \Sigma
\end{aligned}$$

Covariance matrix I

The covariance matrix $C = \{c_{jk}\}$ indicates the tendency of each pair of features (dimensions in a random vector) to vary together, to **co-vary**.

The covariance has several important properties

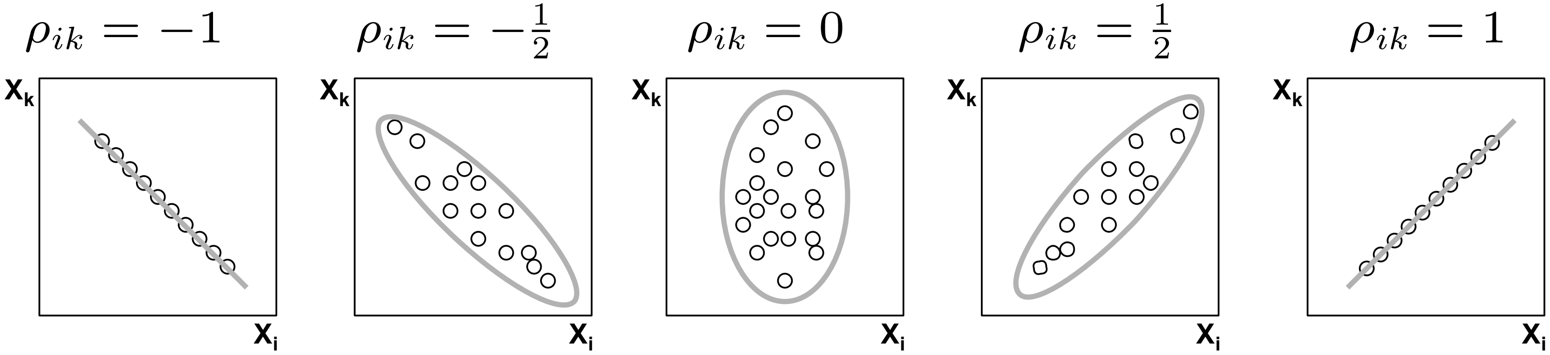
- If X_i and X_k tend to increase together, then $c_{ik} > 0$
- If X_i tends to decrease when X_k increases, then $c_{ik} < 0$
- If X_i and X_k are uncorrelated, then $c_{ik} = 0$

- $|c_{ik}| \leq \sigma_i \, \sigma_k$, where σ_i is the standard deviation of X_i
- $c_{ii} = \sigma_i^2 = \mathrm{Var}\,[X_i]$.

Covariance terms can be expressed as

$$c_{ii} = \sigma_i^2 \quad \text{and} \quad c_{ik} = \rho_{ik}\sigma_i\sigma_k$$

where ρ_{ik} is called the **correlation coefficient**.



Covariance matrix II

The covariance matrix can be reformulated as

$$\Sigma = \mathrm{E}\left[(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T\right] = \mathrm{E}\left[\mathbf{X}\mathbf{X}^T\right] - \boldsymbol{\mu}\boldsymbol{\mu}^T = S - \boldsymbol{\mu}\boldsymbol{\mu}^T$$

with

$$S = \mathrm{E}\left[\mathbf{X}\mathbf{X}^T\right] = \begin{pmatrix} \mathrm{E}\left[X_1 X_1\right] & \dots & \mathrm{E}\left[X_1 X_N\right] \\ \vdots & \dots & \vdots \\ \mathrm{E}\left[X_N X_1\right] & \dots & \mathrm{E}\left[X_N X_N\right] \end{pmatrix}$$

S is called the auto-correlation matrix and contains the same amount of information as the covariance matrix.

The covariance matrix can also be expressed as

$$\Sigma = \Gamma R \Gamma = \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_N \end{pmatrix} \begin{pmatrix} 1 & \rho_{12} & \ldots & \rho_{1N} \\ \rho_{12} & 1 & \ldots & \rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1N} & \rho_{2N} & \ldots & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \ldots & 0 \\ 0 & \sigma_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma_N \end{pmatrix}$$

- A convenient formulation since Γ contains the scales of the features and R retains the essential information of the relationship between the features.

- R is the correlation matrix.

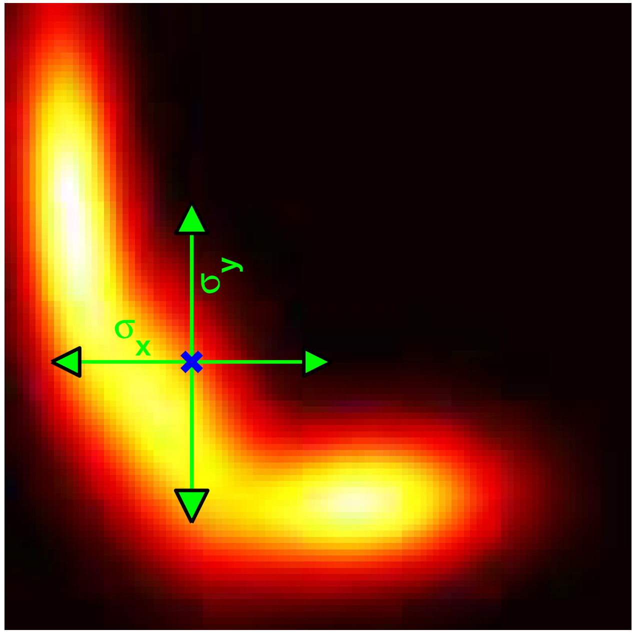
Correlation Vs. Independence

- Two random variables X_i and X_k are uncorrelated if $\mathrm{E}[X_i X_k] = \mathrm{E}[X_i]\,\mathrm{E}[X_k]$. Uncorrelated variables are also called linearly independent.

- Two random variables X_i and X_k are independent if

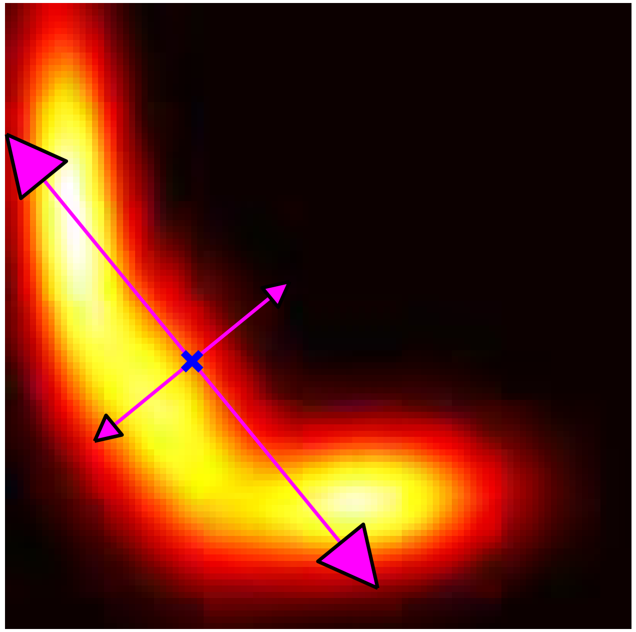
$$p(x_i, x_k) = p(x_i)\, p(x_k) \ \ \forall x_i, x_k$$

Covariance intuition



$$\sigma_x = 0.2249, \quad \sigma_y = 0.2588$$

Covariance intuition



Eigenvectors of Σ are the orthogonal directions where there is the most spread in the underlying distribution.

The eigenvalues indicate the magnitude of the spread.

The Gaussian Distribution

Some Motivation

Where they're used

- Modelling the class conditional pdfs. Frequently, a Gaussian distribution is used or is a building block in modelling it.

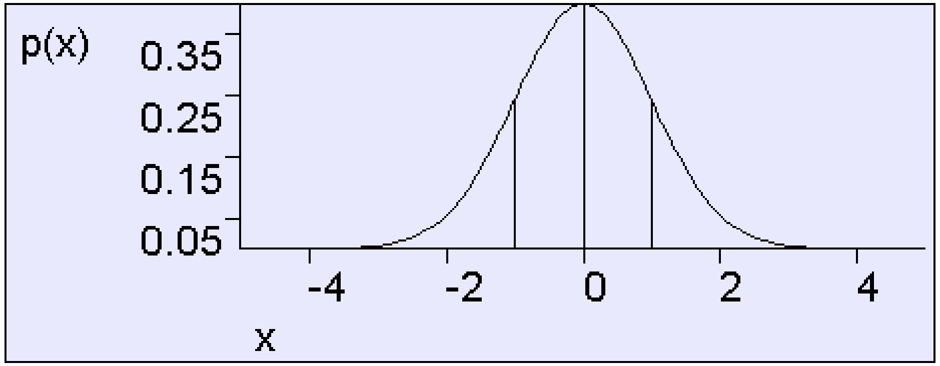
Why they are so important

- They pop up everywhere.
- Need them to understand the optimal Bayes' classifier
- Need them to understand neural networks.
- Need them to understand mixture models.

Unit variance Gaussian

The Gaussian distribution with expected value $\mathrm{E}[X] = 0$ and variance $\mathrm{Var}[X] = 1$ has pdf:

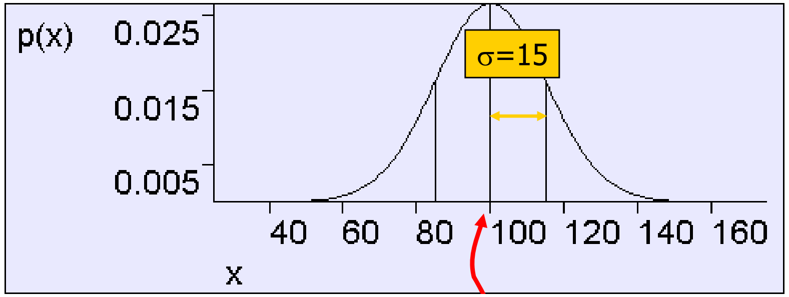
$$p_X(x) = \frac{1}{\sqrt{(2\pi)}} \exp\left(-\frac{x^2}{2}\right)$$



General 1D Gaussian

The Gaussian distribution with expected value $\mathrm{E}[X] = \mu$ and variance $\mathrm{Var}[X] = \sigma^2$ has pdf:

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



Terminology:

Write $X \sim \mathcal{N}(\mu, \sigma^2)$ to denote:

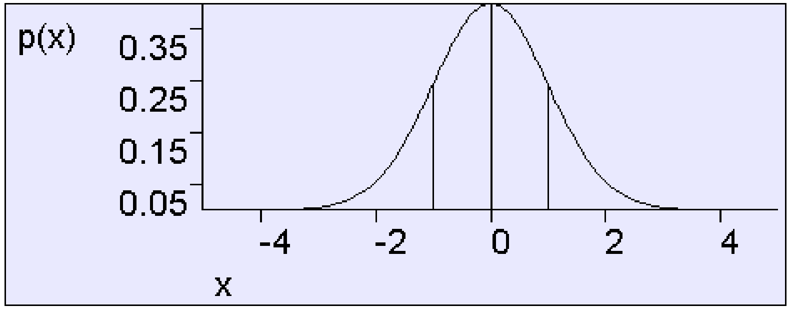
$$\textit{X is distributed as a Gaussian with mean } \mu \textit{ and variance } \sigma^2.$$

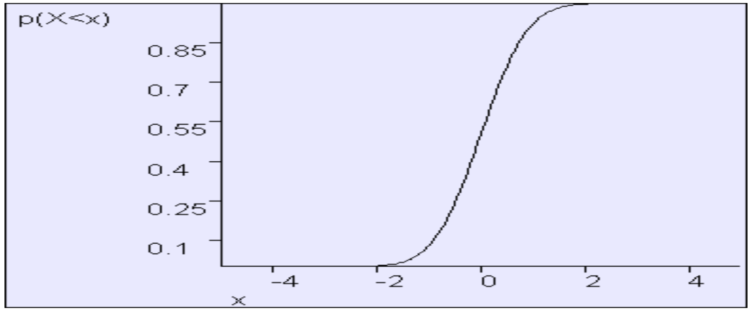
The error function

If $X \sim \mathcal{N}(0,1)$ then erf is defined as

$$\mathrm{erf}(x) = \int_{z=-\infty}^{x} p(z)\,dz = \frac{1}{\sqrt{2\pi}} \int_{z=-\infty}^{x} \exp\left(-\frac{z^2}{2}\right)\,dz$$

Cumulative Distribution of X





Note if $X \sim \mathcal{N}(\mu, \sigma^2)$ then $P(X < x) = \mathrm{erf}\left(\frac{x-\mu}{\sigma}\right)$

The Central Limit Theorem

Assume $X_1, X_2, \ldots, X_N$ are identically and independently distributed (i.i.d.) continuous random variables.

Define

$$z = f(x_1, x_2, \ldots, x_N) = \frac{1}{N} \sum_{i=1}^{N} x_i$$

then

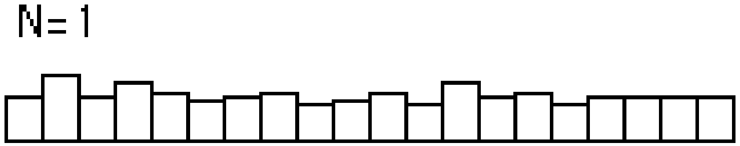
$$p(z) \rightarrow \mathcal{N}(\mathrm{E}\left[X\right], \ \mathrm{Var}\left[X\right]) \text{ as } N \rightarrow \infty.$$

Frequently used as a justification for assuming Gaussian noise.

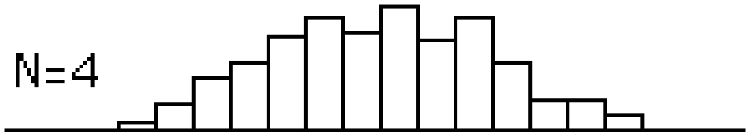
Illustration

500 experiments were performed using a uniform distribution.

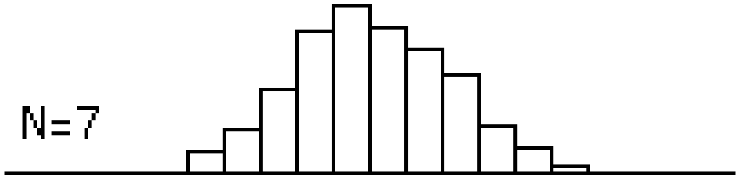
- For $N = 1$, one sample was drawn from the distribution and its mean was recorded (for each of the 500 experiments). The histogram of the result shows a uniform density.



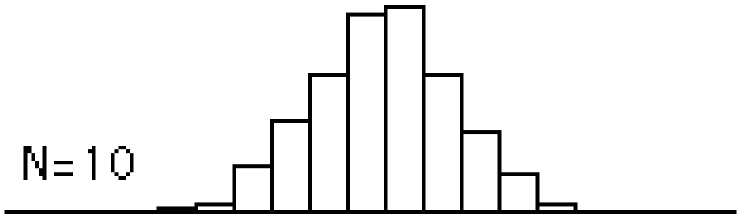
- For $N = 4$, 4 samples were drawn from the distribution and the mean of these 4 samples was recorded (for each of the 500 experiments). The histogram starts to show a Gaussian shape.



- Similarly for $N = 7$



- Similarly for $N = 10$



As N grows the histogram increasingly resembles a Gaussian.

Bivariate Gaussian

Write random variable $\mathbf{X} = \begin{pmatrix} X \\ Y \end{pmatrix}$. Define $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ to mean

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

where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix} \qquad \Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}$$

and Σ must be symmetric and non-negative definite.

It turns out for a Gaussian distribution $\mathrm{E}\,[\mathbf{X}] = \boldsymbol{\mu}$ and $\mathrm{Cov}\,[\mathbf{X}] = \Sigma$.

General Gaussian

Have a random variable $\mathbf{X} = \left(X_1, X_2, \ldots, X_m\right)^T$.

Then define $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ as

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{m}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)$$

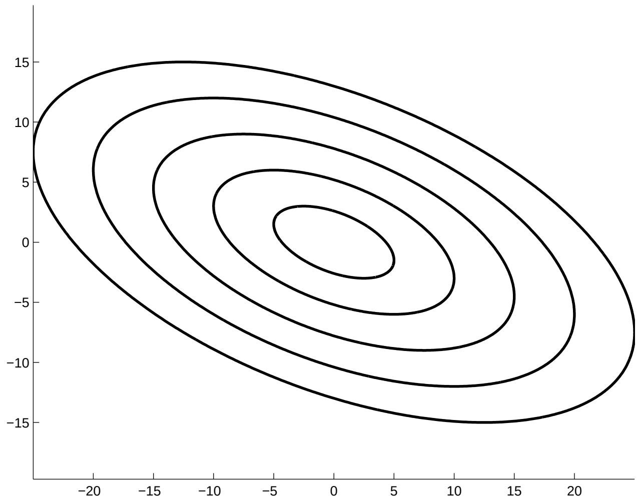
where

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$

and Σ must be symmetric non-negative definite.

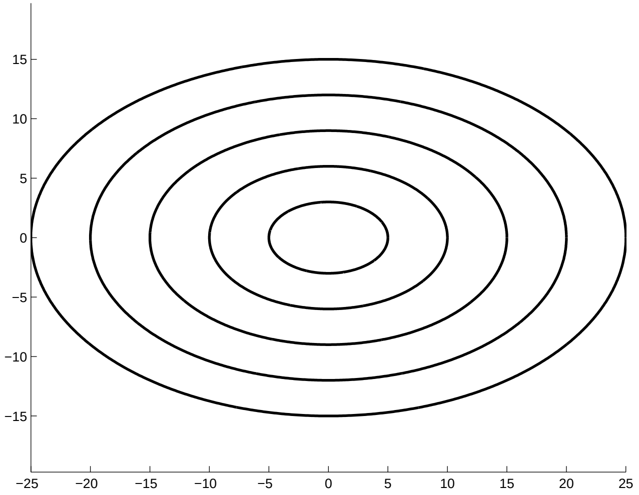
General Gaussian

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \cdots & \sigma_m^2 \end{pmatrix}$$



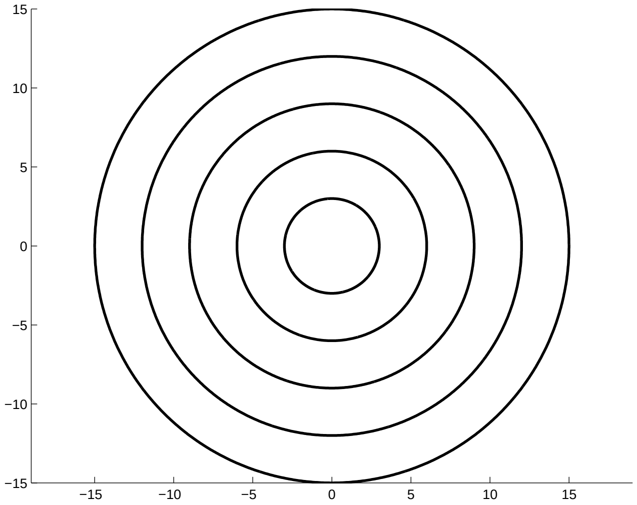
Axis-aligned Gaussian

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma_2^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_{m-1}^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma_m^2 \end{pmatrix}$$



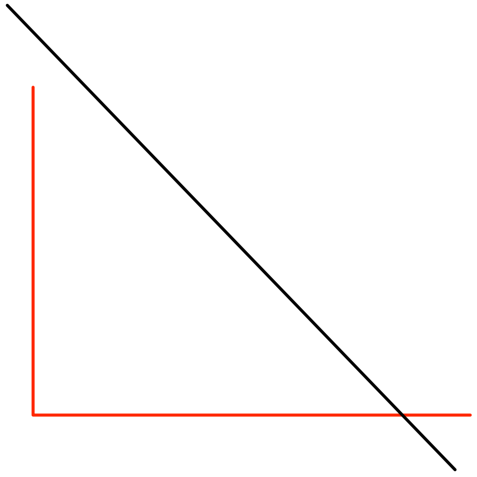
Spherical Gaussian

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & \cdots & 0 & \sigma^2 \end{pmatrix}$$



Degenerate Gaussian

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_m \end{pmatrix}, \quad |\Sigma| = 0$$



Recap

- Have seen the formulae for Gaussians
- You should have an intuition of how they behave
- Have some confidence in $reading$ a Gaussian's covariance matrix.

How can we transform a random variable with a non-diagonal covariance matrix to one with a diagonal covariance matrix?

Eigenvectors and eigenvalues

Given an $m \times m$ matrix A.

Definition:

$\mathbf{v}$ an eigenvector of A if there exists a scalar λ (eigenvalue) such that

$$A\mathbf{v} = \lambda\mathbf{v}$$

How to compute them:

For an eigenvector-eigenvalue pair of A:

$$A\mathbf{v} = \lambda\mathbf{v} \;\Rightarrow\; A\mathbf{v} - \lambda\mathbf{v} = \mathbf{0}$$

$$\Rightarrow\; (A - \lambda I)\mathbf{v} = \mathbf{0} \;\Rightarrow\; \begin{cases} \mathbf{v} = \mathbf{0} & \text{trivial soln} \\ (A - \lambda I) & \text{is rank deficient} \end{cases}$$

For the non-trivial solution when $(A - \lambda I)$ being rank deficient implies

$$\det(A - \lambda I) = 0 \;\Rightarrow\; \underbrace{\lambda^m + a_1\lambda^{m-1} + \cdots + a_{m-1}\lambda + a_m}_{\textbf{characteristic equation}} = 0$$

Solve this **characteristic equation** to obtain possible values for λ and given those then compute their corresponding eigenvector.

Some terminology:

Matrix formed by the column eigenvectors of A is called the modal matrix M:

$$M = \begin{pmatrix} \uparrow & \uparrow & \uparrow & & \uparrow \\ \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \ldots & \mathbf{v}_m \\ \downarrow & \downarrow & \downarrow & & \downarrow \end{pmatrix}$$

Let Λ be the diagonal matrix with A's eigenvalues on the main diagonal:

$$\Lambda = \begin{pmatrix} \lambda_1 & & & & \\ & \lambda_2 & & & \\ & & \lambda_3 & & \\ & & & \ddots & \\ & & & & \lambda_m \end{pmatrix}$$

Matrix Λ is the canonical form of A.

Properties of A and implications for its eigenvalues:

A non-singular	$\Rightarrow$	All eigenvalues are non-zero.
A real and symmetric	$\Rightarrow$	All eigenvalues are real **and** Eigenvectors associated with distinct eigenvalues are orthogonal.
A positive definite	$\Rightarrow$	All eigenvalues are positive.

Eigen-decomposition of A

If A has non-degenerate eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_m$ (no two λ_i, λ_j have the same value) and M is A's modal matrix then

$$\begin{aligned} AM &= A[\mathbf{v}_1\, \mathbf{v}_2\, \ldots\, \mathbf{v}_m] \\ &= [A\mathbf{v}_1\, A\mathbf{v}_2\, \ldots\, A\mathbf{v}_m] \\ &= [\lambda_1\mathbf{v}_1\, \lambda_2\mathbf{v}_2\, \ldots\, \lambda_m\mathbf{v}_m] = M\Lambda \\ \Longrightarrow \quad A &= M\Lambda M^{-1} \leftarrow \text{ the eigen-decomposition of } A \end{aligned}$$

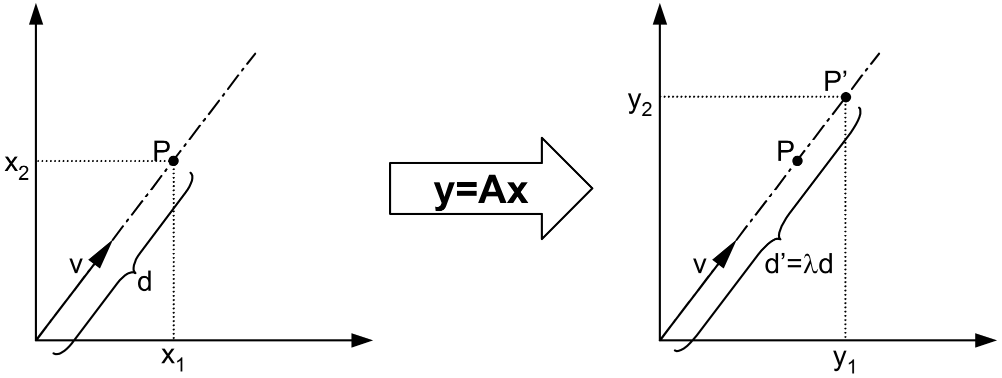
Consequently:

$$\begin{aligned} A^2 &= M\Lambda^2 M^{-1}, \\ A^n &= M\Lambda^n M^{-1} \quad \text{for } n = 1, 2, 3, \ldots \\ A^{-1} &= (M\Lambda M^{-1})^{-1} = M\Lambda^{-1} M^{-1} \end{aligned}$$

Interpretation

View matrix A as a linear transformation:

An eigenvector $\mathbf{v}$ represents an invariant direction of the transformation.

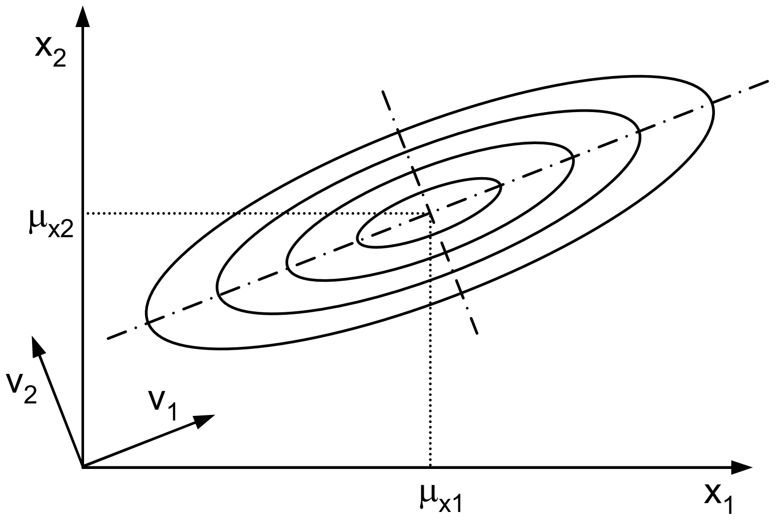


When transformed by A, any point lying on the direction defined by $\mathbf{v}$ will remain on that direction, and its magnitude will be multiplied by the corresponding eigenvalue λ.

Back to Covariance Matrices

Given Σ the covariance of a Gaussian distribution

Eigenvectors of Σ are the principal directions of the distribution.



Their eigenvalues are the variances in these principal direction.

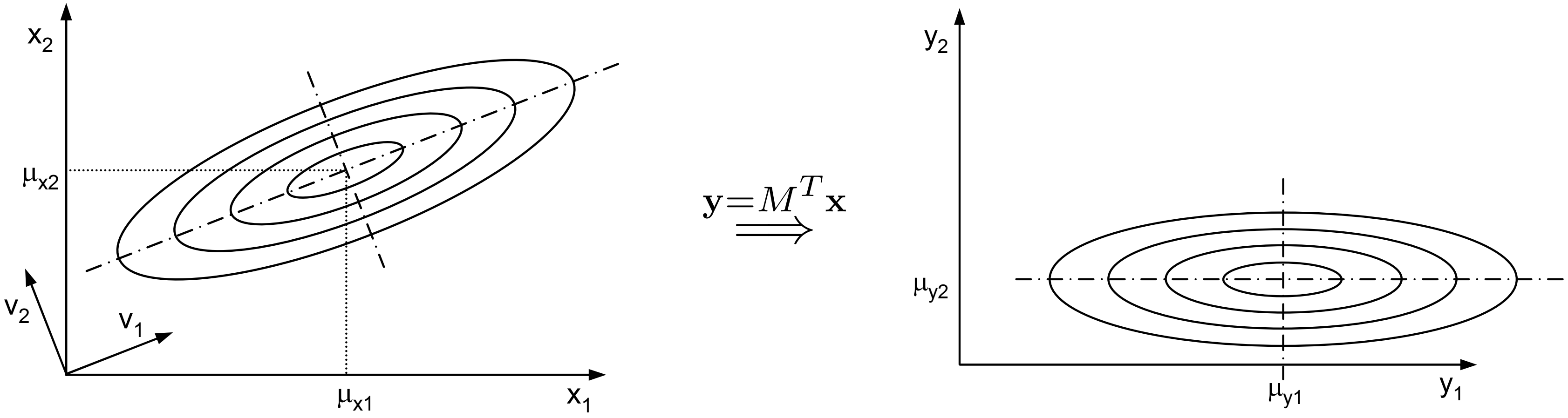
Linear transformation defined by the eigenvectors of Σ

Obtain uncorrelated vector regardless of the form of the distribution.

Let M be the modal matrix of the covariance matrix $\Sigma \implies \Sigma M = M\Lambda$

Define the transformation: $\mathbf{y} = M^T\mathbf{x}$

if $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ then $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \Lambda)$ as Σ is a symmetric matrix with real values. This in turn means $\mathbf{y}$ is a Gaussian random variable with diagonal covariance matrix.



Manipulations of Normally Distributed random variables

Linear transforms remain Gaussian

Assume $\mathbf{X}$ is an $m-$dimensional Gaussian random variable

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

Define $\mathbf{Y}$ to be a $p-$dimensional random variable with

$$\mathbf{Y} = A\,\mathbf{X}$$

where A is a $p \times m$ matrix. Then

$$\mathbf{Y} \sim \mathcal{N}(A\boldsymbol{\mu}, A\Sigma A^T)$$

Gaussian marginals are Gaussian

Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)^T$ be a multivariate Gaussian random variable and define subsets of its variables as follows

$$\mathbf{X} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \quad \text{with} \quad \mathbf{U} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} X_{p+1} \\ \vdots \\ X_m \end{pmatrix}$$

If

$$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_u \\ \boldsymbol{\mu}_v \end{pmatrix}, \begin{pmatrix} \Sigma_{uu} & \Sigma_{uv} \\ \Sigma_{uv}^T & \Sigma_{vv} \end{pmatrix} \right)$$

Then $\mathbf{U}$ is also distributed as a Gaussian with $\mathbf{U} \sim \mathcal{N}(\boldsymbol{\mu}_u, \Sigma_{uu})$.

How do we prove this using the result on the previous slide?

Adding 2 independent Gaussians

If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{X}})$ and $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{Y}}, \Sigma_{\mathbf{Y}})$ and $\mathbf{X} \perp\!\!\!\perp \mathbf{Y}$ then

$$\mathbf{X} + \mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{X}} + \boldsymbol{\mu}_{\mathbf{Y}}, \Sigma_{\mathbf{X}} + \Sigma_{\mathbf{Y}})$$

If $\mathbf{X}$ and $\mathbf{Y}$ are not independent but uncorrelated the above result **does not** hold.

Conditional of a Gaussian

Assume that

$$\begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\boldsymbol{\mu}_u\\ \boldsymbol{\mu}_v\end{pmatrix}, \begin{pmatrix}\Sigma_{uu} & \Sigma_{uv}\\ \Sigma_{uv}^T & \Sigma_{vv}\end{pmatrix}\right)$$

then $\mathbf{U} \,|\, \mathbf{V} \sim \mathcal{N}\left(\boldsymbol{\mu}_{u|v}, \Sigma_{u|v}\right)$ where

$$\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \Sigma_{uv}^T\,\Sigma_{vv}^{-1}(\mathbf{V} - \boldsymbol{\mu}_v)$$

$$\Sigma_{u|v} = \Sigma_{uu} - \Sigma_{uv}^T\,\Sigma_{vv}^{-1}\,\Sigma_{uv}$$

Conditional of a Gaussian

Consider the marginal mean:

$$\boldsymbol{\mu}_{u|v} = \boldsymbol{\mu}_u + \Sigma_{uv}^T \, \Sigma_{vv}^{-1} (\mathbf{V} - \boldsymbol{\mu}_v)$$

When $\mathbf{V} = \boldsymbol{\mu}_v \Rightarrow$ the conditional mean of $\mathbf{U}$ is $\boldsymbol{\mu}_u$

Marginal mean is a linear function of $\mathbf{V}$.

Consider the conditional covariance:

$$\Sigma_{u|v} = \Sigma_{uu} - \Sigma_{uv}^T \, \Sigma_{vv}^{-1} \, \Sigma_{uv}$$

Conditional variance is $\leq$ than the marginal variance.

Conditional varaince is independent of the given value of $\mathbf{V}$.

Gaussians and the chain rule

Let A be a constant matrix if

$$\mathbf{U} \,|\, \mathbf{V} \sim \mathcal{N}(A\mathbf{V}, \Sigma_{u|v}) \quad \textbf{and} \quad \mathbf{V} \sim \mathcal{N}(\boldsymbol{\mu}_v, \Sigma_{vv})$$

then

$$\begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

with

$$\boldsymbol{\mu} = \begin{pmatrix} A\boldsymbol{\mu}_v \\ \boldsymbol{\mu}_v \end{pmatrix} \quad \textbf{and} \quad \Sigma = \begin{pmatrix} A\Sigma_{vv}A^T + \Sigma_{u|v} & A\Sigma_{vv} \\ (A\Sigma_{vv})^T & \Sigma_{vv} \end{pmatrix}$$

Are these manipulations useful?

Bayesian inference

Consider the following problem:

In the world as a whole, IQs are drawn from a Gaussian $\mathcal{N}(100, 15^2)$.

If you take an IQ test you'll get a score that, on average (over many tests) will be your IQ. Because of noise on any one test the score will often be a few points lower or higher than your true IQ. Thus assume we have the conditional distribution

$$\text{Score} \,|\, \text{IQ} \sim \mathcal{N}(IQ, 10^2)$$

If you take the IQ test and get a score of 130 what is the most likely value of your IQ given this piece of evidence?

IQ Example

Which distribution should we calculate?

How can we get an expression for this distribution from the distributions of Score | IQ and IQ and the manipulations we have described?

Plan

This we know:

$$\text{IQ} \sim \mathcal{N}(100, 15^2), \quad \text{Score} \,|\, \text{IQ} \sim \mathcal{N}(\text{IQ}, 15^2), \quad \text{Score} = 130$$

Want to find the distribution $p(\text{IQ} \,|\, \text{Score} = 130)$.

Plan:

- Use the chain rule to compute the distribution of $\begin{pmatrix} \text{Score} \\ \text{IQ} \end{pmatrix}$ from distributions of IQ and Score | IQ
- Swap the order of the random variables to get $\begin{pmatrix} \text{IQ} \\ \text{Score} \end{pmatrix}$'s distribution

- From $\begin{pmatrix} \text{IQ} \\ \text{Score} \end{pmatrix}$ compute the conditional distribution to get the distribution of $\text{IQ} \mid (\text{Score} = 130)$

What is the best estimate for test taker's IQ?

Today’s assignment

Pen & Paper assignment

- Details available on the course website.
- You will be asked to perform some simple Bayesian reasoning.
- Mail me about any errors you spot in the `Exercise` notes.
- I will notify the class about errors spotted and corrections via the course website and mailing list.

Consider this..

Mrs S. is found stabbed in the garden of her family home in the USA. Her husband Mr. S. behaves strangely after her death and is considered as a suspect.

On investigation the police discover that Mr S. had beaten up his wife on at least nine previous occasions. The prosecution offers up this information as evidence in favour of the hypothesis that Mr S is guilty of the murder.

However, Mr S.'s lawyer disputes this by saying "*statistically, only one in a thousand wife-beaters actually goes on to murder his wife. So the wife beating is not strong evidence at all. In fact given the wife-beating evidence alone, it's extremely unlikely that he would be the murderer of his wife - only a 1/1000 chance. You should therefore find him innocent.*"

The prosecution replies with these two following empirical facts. In the USA it is estimated that 2 million woman are abused each year by their partners. (Let's assume there are 100 million adult women in the USA). In 1994, 4739 women were victims of homicide, of those, 1326 women (28%) were killed by husbands and boyfriends.

Question:

Is the lawyer right to imply that the history of wife-beating does not point to Mr S.'s being the murderer? Or is the lawyer's reasoning flawed? If the latter, what is wrong with his argument?