Undirected network reconstruction – part 1

Wessel N. van Wieringen
w.n.van.wieringen@vu.nl

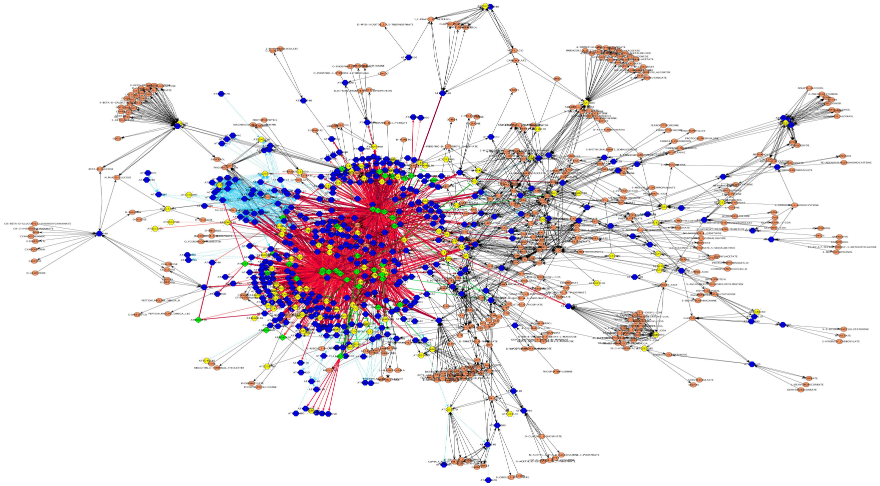
Department of Epidemiology and Biostatistics, VUmc
& Department of Mathematics, VU University
Amsterdam, The Netherlands

What?

Molecular biology aims to understand the molecular processes that occur in the cell. That is, e.g.:
→ which molecules present in the cell interact?
→ how is this coordinated?

For many cellular processes, it is unknown which genes play what role.

Goal
Reconstruct the cellular regulatory network.

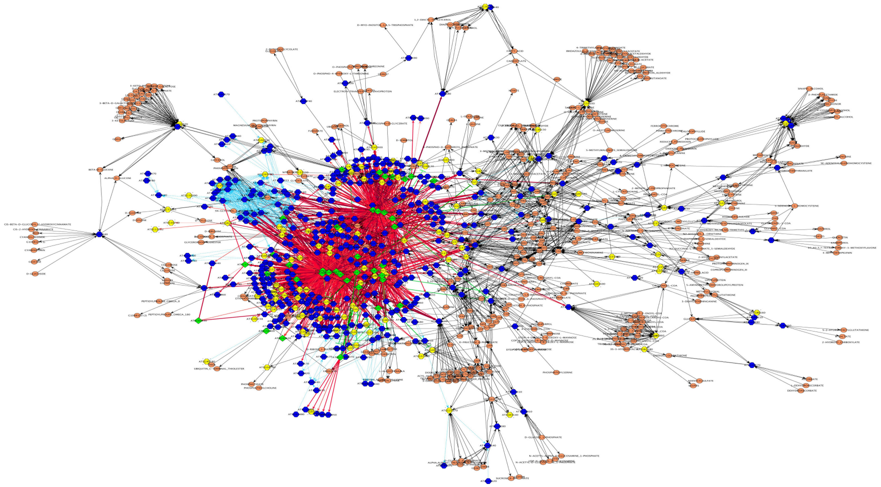


Why?

Negative motivation

→ Differentially expressed genes: boring!

→ Yet another clustering?

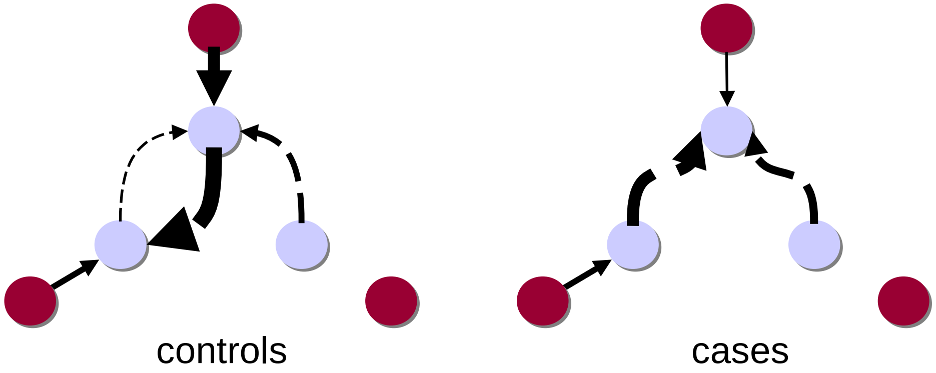


Positive motivation

→ Fancy plot.

→ Different insight.

→ Network medicine
(e.g. biomarker:
gene-gene interaction)



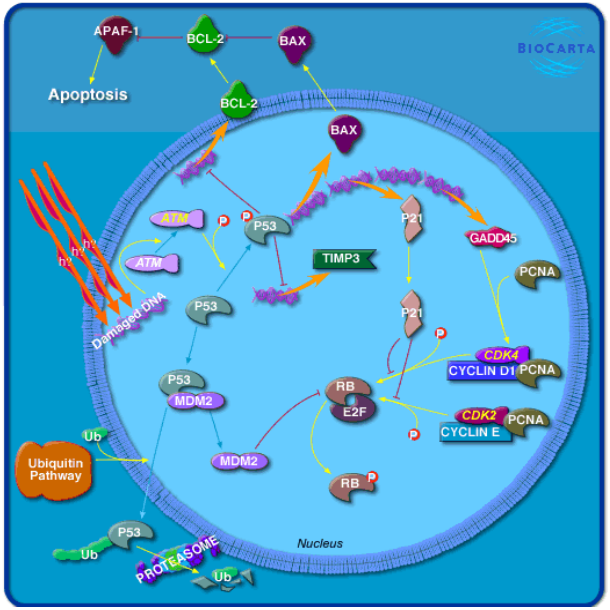
Pathway = network

Pathway = chain of chemical reactions (that processes a signal)

≈ a set of genes believed to carry out one function

Pathways are loosely defined using repositories, such as:

- KEGG
- BioCarta
- GennMapp
- Reactome
- GO
- String

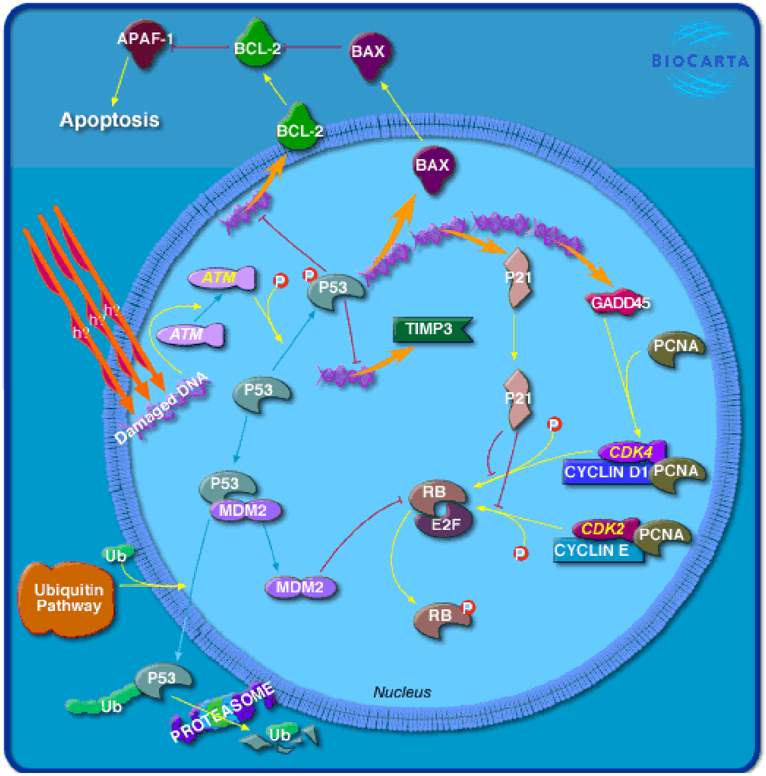


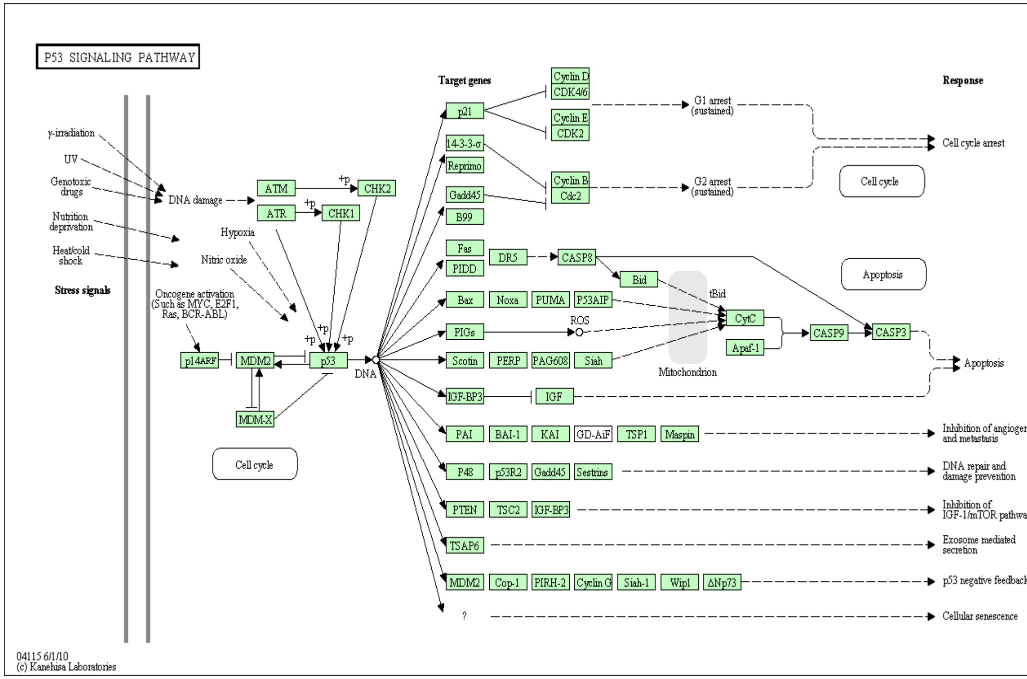
BioCarta: p53 signalling pathway

Pathway = network

BioCarta
p53 signalling pathway

KEGG
p53 signalling pathway





http://www.biocarta.com/pathfiles/h_p53Pathway.asp

http://www.genome.jp/kegg/pathway/hsa/hsa04115.html

How?

Download from repository

→ Which? Reliable?

→ Knowledge is incomplete and biased towards a few well-studied pathways.

→ Does it apply to the situation at hand?

Reconstruct from data

→ Data is a rare and valuable commodity!

Synthesis

Reconstruct from data with the repository as a suggestion

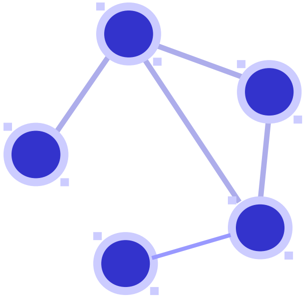
Network

Pathways are represented by a *graph* or *network*.

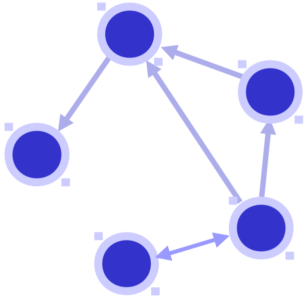
● *node* or *vertex*, representing a gene.

→ ↔ *edge* or *arrow*, representing an interaction between two genes.

— *undirected* and *directed* edges (≈ "association")



undirected
(focus here)



directed

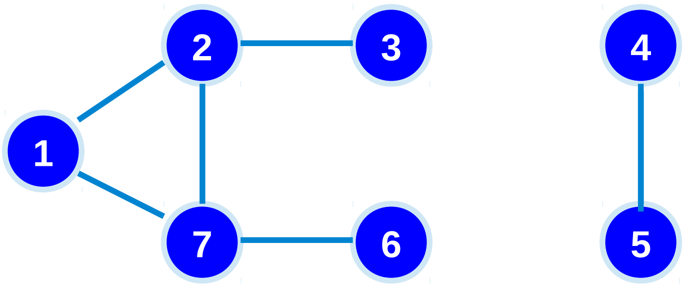
Network

Edge operationalization = direct relation
(Formally: conditional dependence)

Direct relation
Relation between two nodes without mediation of other nodes.

Indirect relation
Relation between two nodes through mediating other nodes.

No relation
None of the above.



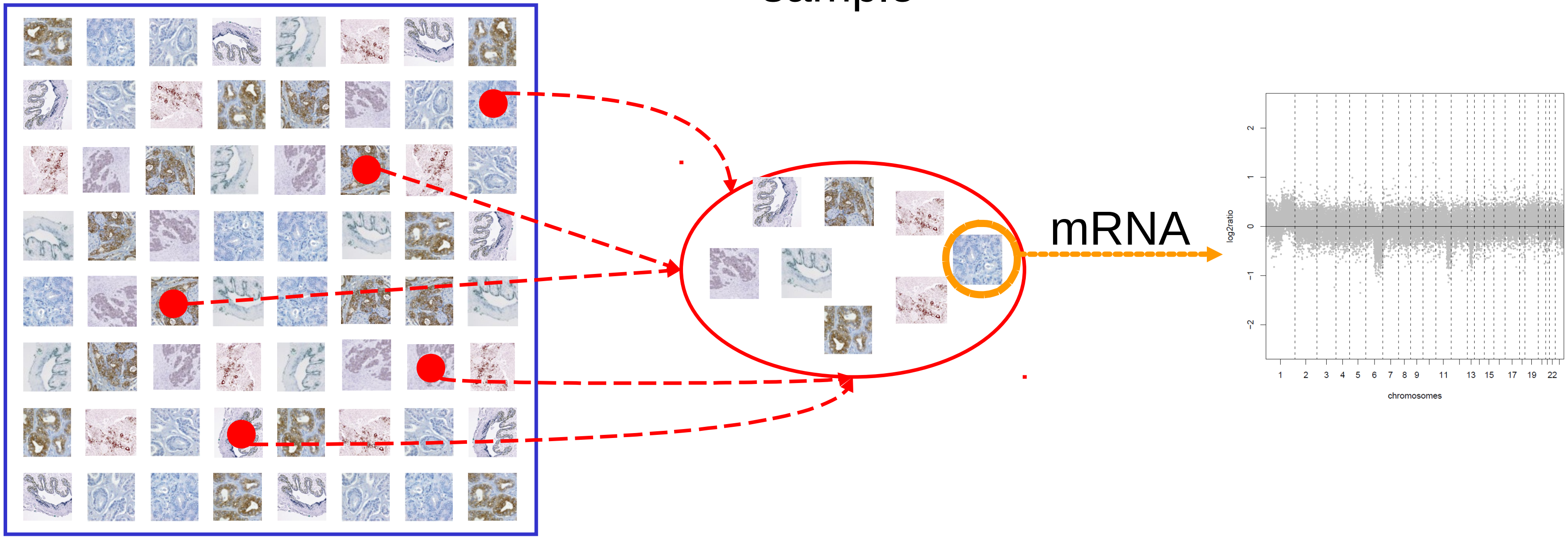
Relations
→ node 1 and 2: directly
→ node 3 and 6: indirectly
→ node 4 and 7: none
→ ...

With?

cancer tissue bank

random sample

profiling



mRNA

- in vivo
- cross-sectional
- time-course

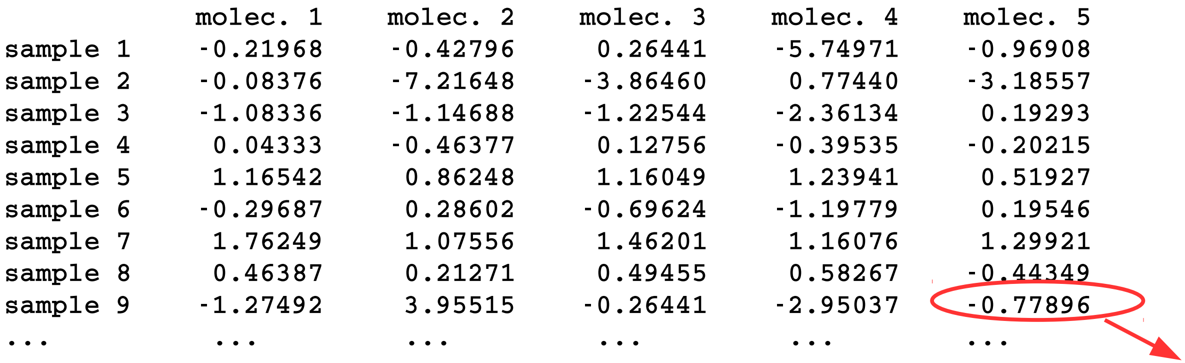
With?

Data

Available to reconstruct which molecular interact:

→ molecular profiles of *n* samples,

→ each profile comprises *p* features.



	molec. 1	molec. 2	molec. 3	molec. 4	molec. 5
sample 1	-0.21968	-0.42796	0.26441	-5.74971	-0.96908
sample 2	-0.08376	-7.21648	-3.86460	0.77440	-3.18557
sample 3	-1.08336	-1.14688	-1.22544	-2.36134	0.19293
sample 4	0.04333	-0.46377	0.12756	-0.39535	-0.20215
sample 5	1.16542	0.86248	1.16049	1.23941	0.51927
sample 6	-0.29687	0.28602	-0.69624	-1.19779	0.19546
sample 7	1.76249	1.07556	1.46201	1.16076	1.29921
sample 8	0.46387	0.21271	0.49455	0.58267	-0.44349
sample 9	-1.27492	3.95515	-0.26441	-2.95037	-0.77896
...	...	...	...	...	...

≈ activity

Repository

Prior knowledge on network

Gaussian graphical model

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

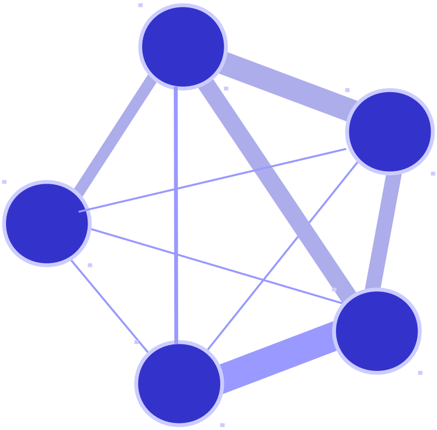
How?

Roadmap

data

```
         sample 1    sample 2    ...
gene 1   -0.21968    -0.42796    ...
gene 2   -0.08376    -7.21648    ...
gene 3   -1.08336    -1.14688    ...
gene 4    0.04333    -0.46377    ...
...     ...   ...    ...
```

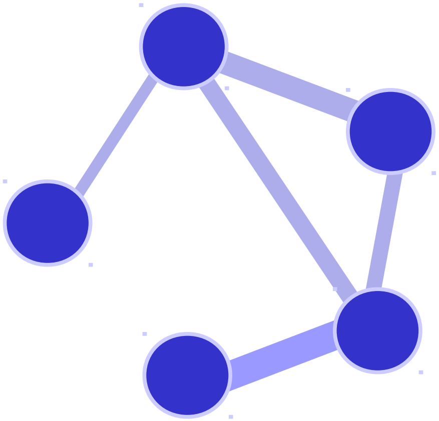
edge strength measure



statistical test

edge strength significantly different from zero: *edge*!

inferred network



(Conditional) independence graph

Whittaker (1990), Chapter 3.

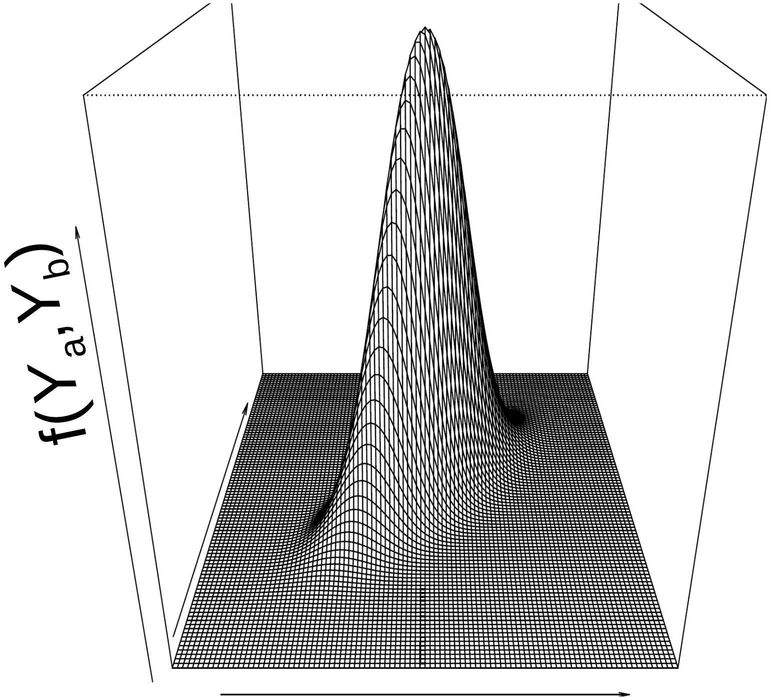
CIG

A *p*-variate random variable $\mathbf{Y}$ is a vector of *p* univariate random variables.

These univariate random variables are considered together when they may be related in some sense.

A *joint density* of a *p-variate* random variable $\mathbf{Y}$ specifies the (relative) probability of observing a particular realization of $\mathbf{Y}$.

joint density of Y_a and Y_b



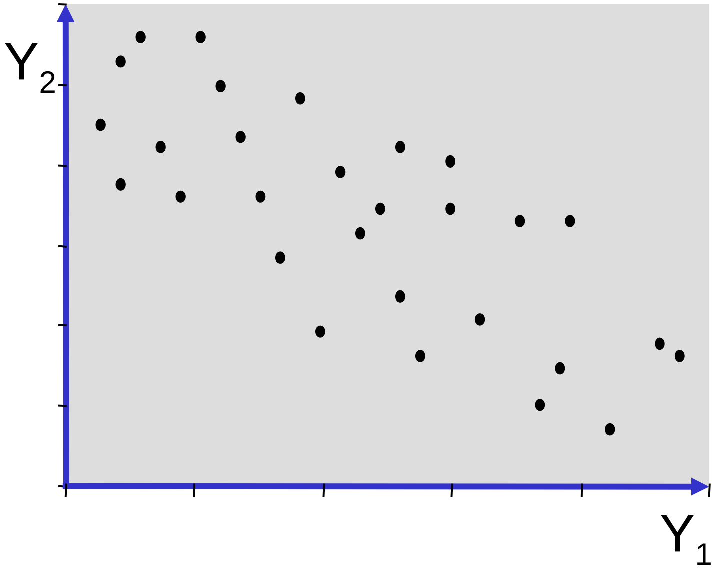
(Conditional) independence

Joint density (example)

The joint density of tossing of two coins:

		coin 1	
		Head	Tail
coin 2	Head	1/4	1/4
	Tail	1/4	1/4

The joint density of the expression levels of two genes describes how their data are distributed in the 2-dim plane:



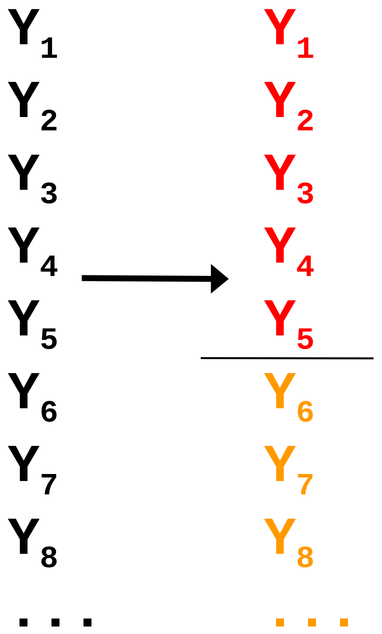
(Conditional) independence

Consider a p-variate random variable $\mathbf{Y}$.

Suppose the p variates can be divided into two exhaustive and mutually exclusive subsets A and B, i.e.:

$$\begin{array}{ll} i) & A, B \subset \{1, 2, \ldots, p\} \\ ii) & A \cap B = \emptyset \\ iii) & A \cup B = \{1, 2, \ldots, p\} \end{array}$$

Let $\mathbf{Y}_a$ and $\mathbf{Y}_b$ be random vectors obtained by restricting $\mathbf{Y}$ to only those variates that correspond to the elements of subset A and B, resp..



Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7 Y_8 . . .

Y_1 Y_2 Y_3 Y_4 Y_5 Y_6 Y_7 Y_8 . . .

(Conditional) independence

The random variables $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are *independent* if and only if the joint probability density function $f_{\mathbf{Y}_a,\mathbf{Y}_b}$ satisfies:

$$f_{\mathbf{Y}_a,\mathbf{Y}_b}(\mathbf{y}_a,\mathbf{y}_b) \quad = \quad f_{\mathbf{Y}_a}(\mathbf{y}_a)\, f_{\mathbf{Y}_b}(\mathbf{y}_b)$$

for all values of $\mathbf{y}_a$ and $\mathbf{y}_b$.

Hence, under independence the joint density factorizes into the product of the marginal densities, e.g.:

$$f_{\mathbf{Y}_a}(\mathbf{y}_a) \quad = \quad \int f_{\mathbf{Y}_a,\mathbf{Y}_b}(\mathbf{y}_a,\mathbf{y}_b)\, d\mathbf{y}_b$$

Independence between $\mathbf{Y}_a$ and $\mathbf{Y}_b$ is denoted by:

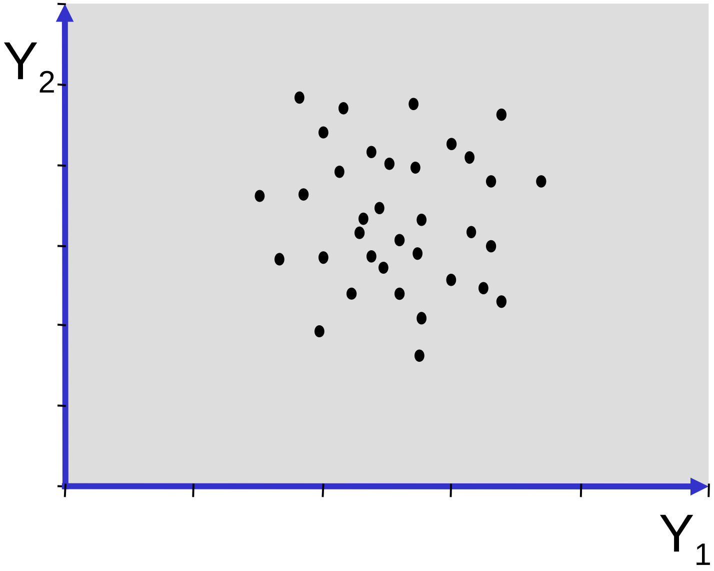
$$\mathbf{Y}_a \perp\!\!\!\perp \mathbf{Y}_b$$

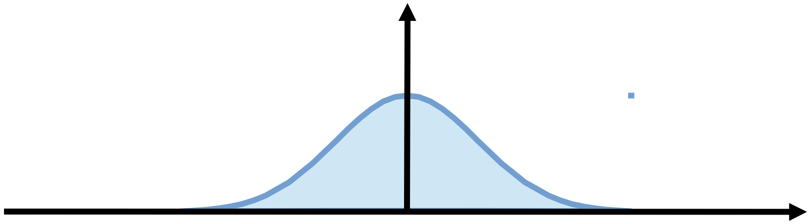
(Conditional) independence

Example

How about two genes?

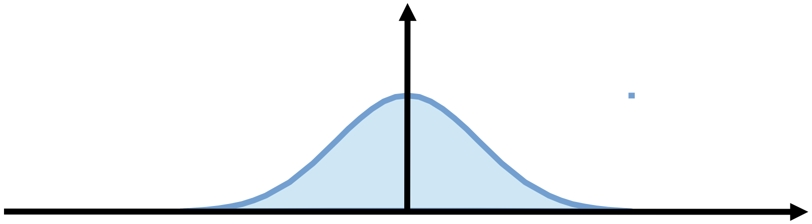
If knowledge of Y_a, the expression level of gene A, does *not* affect the (distribution of the) expression levels of gene B, the two genes are said to be *independent*.





distribution of Y_b when Y_a = 2

=



distribution of Y_b when Y_a = -3

(Conditional) independence

An equivalent definition
The random variables $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are *independent* if and only if:

$$f_{\mathbf{Y}_a|\mathbf{Y}_b}(\mathbf{y}_a, \mathbf{y}_b) \quad = \quad f_{\mathbf{Y}_a}(\mathbf{y}_a)$$

for all values of $\mathbf{y}_a$ and $\mathbf{y}_b$. Hence, the conditional and marginal densities are identical.

This follows from:

$$\begin{aligned} f_{\mathbf{Y}_a|\mathbf{Y}_b}(\mathbf{y}_a, \mathbf{y}_b) \quad &= \quad f_{\mathbf{Y}_a,\mathbf{Y}_b}(\mathbf{y}_a, \mathbf{y}_b) \, / \, f_{\mathbf{Y}_b}(\mathbf{y}_b) \\ &= \quad f_{\mathbf{Y}_a}(\mathbf{y}_a) \, f_{\mathbf{Y}_b}(\mathbf{y}_b) \, / \, f_{\mathbf{Y}_b}(\mathbf{y}_b) \\ &= \quad f_{\mathbf{Y}_a}(\mathbf{y}_a) \end{aligned}$$

(Conditional) independence

Consider a pathway comprising of two genes.

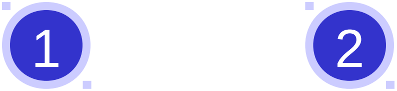
Expression levels of genes 1 and 2 are *independent*:

$$Y_1 \perp\!\!\!\perp Y_2$$

Hence:

$$f_{Y_1,Y_2}(y_1, y_2) = f_{Y_1}(y_1)\, f_{Y_2}(y_2)$$

Graph:



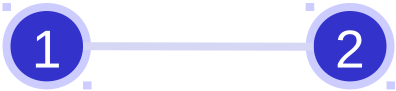
Expression levels of genes 1 and 2 are *dependent*:

$$Y_1 \not\!\perp\!\!\!\perp Y_2$$

Hence:

$$f_{Y_1,Y_2}(y_1, y_2) \neq f_{Y_1}(y_1)\, f_{Y_2}(y_2)$$

Graph:



(Conditional) independence

Consider a pathway comprising of two genes.

Question

The density of the expression levels of genes 1 and 2 is:

$$f_{Y_1,Y_2}(y_1, y_2) \quad = \quad C \exp(-y_1^2 - 3y_2^2)$$

with *C* a suitable constant. Are Y_1 and Y_2 in- or dependent?

Question

The density of the expression levels of genes 1 and 2 is:

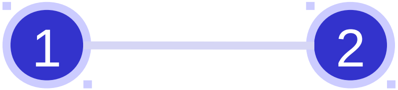
$$f_{Y_1,Y_2}(y_1, y_2) \quad = \quad C \exp(-y_1^2 - 3y_2^2 + 2y_1 y_2)$$

with *C* a suitable constant. Are Y_1 and Y_2 in- or dependent?

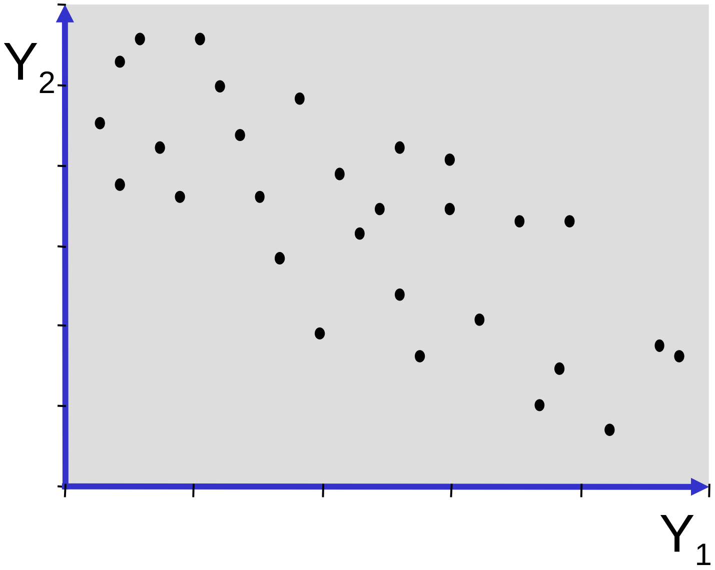
(Conditional) independence

Consider a pathway comprising of two genes.

Expression levels of genes 1 and 2 are *dependent*: $Y_1 \not\perp Y_2$

Graph: 

Data:



If a low value for *Y_1* is observed, it is more likely to observe a high value for *Y_2*.

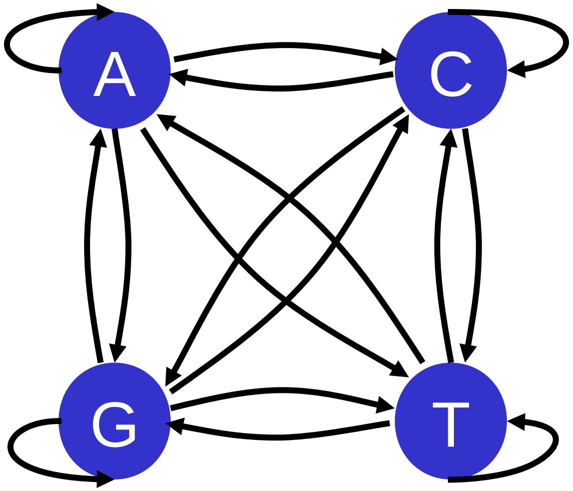
And vice versa.

(Conditional) independence

Recall (from the first 4 lectures)

DNA sequence modeled by a 1st order Markov chain:

$$\mathbf{P} = \begin{array}{cc} & \overbrace{\begin{array}{cccc} A & C & G & T \end{array}}^{\textit{to}} \\ \begin{array}{c} \text{f}\\ \text{r}\\ \text{o}\\ \text{m} \end{array}\left\{\begin{array}{c} A\\ C\\ G\\ T \end{array}\right. & \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.7 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.5 & 0.1 \end{pmatrix} \end{array}$$



Question

Are X_t and X_{t+1} independent? What about X_t and X_{t+2}?

(Conditional) independence

Consider a *p*-variate random variable $\mathbf{Y}$.

Suppose the *p* variates can be divided into three exhaustive and mutually exclusive subsets A ,B, and C i.e.:

$$
\begin{array}{ll}
i) & A, B, C \subset \{1, 2, \ldots, p\} \\
ii) & A \cap B = \emptyset, A \cap C = \emptyset, B \cap C = \emptyset \\
iii) & A \cup B \cup C = \{1, 2, \ldots, p\}
\end{array}
$$

Denote by $\mathbf{Y}_a$, $\mathbf{Y}_b$ and $\mathbf{Y}_c$ the random vectors that are obtained by restricting $\mathbf{Y}$ to only those variates that correspond to the elements of subset A, B and C, respectively.

(Conditional) independence

The random variables $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are *conditional independent* on $\mathbf{Y}_c$f and only if:

$$f_{\mathbf{Y}_a,\mathbf{Y}_b\,|\,\mathbf{Y}_c}(\mathbf{y}_a,\mathbf{y}_b,\mathbf{y}_c) \\ = \quad f_{\mathbf{Y}_a\,|\,\mathbf{Y}_c}(\mathbf{y}_a,\mathbf{y}_c)\,f_{\mathbf{Y}_b\,|\,\mathbf{Y}_c}(\mathbf{y}_b,\mathbf{y}_c)$$

for all values of $\mathbf{y}_a$, $\mathbf{y}_b$ and $\mathbf{y}_c$.

This conditional independence is denoted as:

$$(\mathbf{Y}_a\,|\,\mathbf{Y}_c) \perp\!\!\!\perp (\mathbf{Y}_b\,|\,\mathbf{Y}_c)$$

Or, more commonly:

$$\mathbf{Y}_a \perp\!\!\!\perp \mathbf{Y}_b\,|\,\mathbf{Y}_c$$

(Conditional) independence

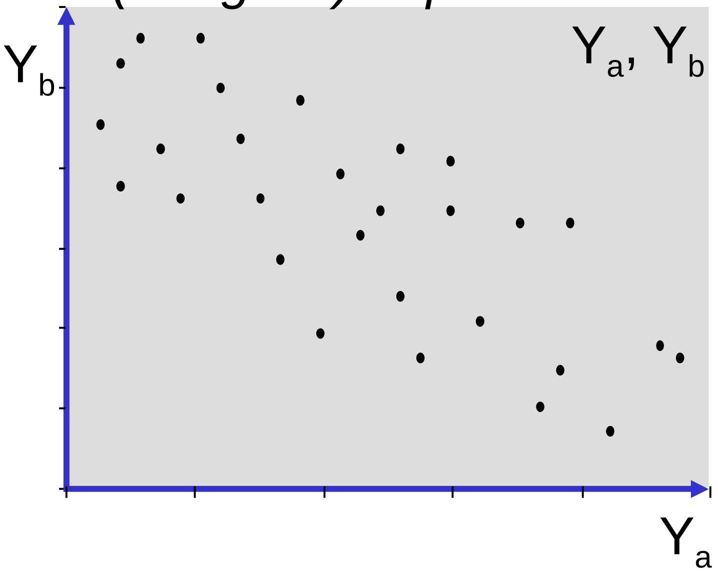
Example

Consider:

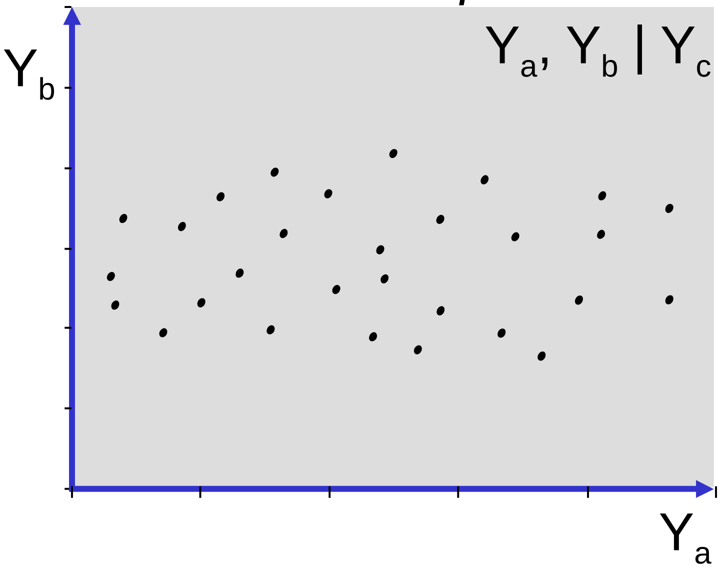
$$\begin{cases} Y_a = Y_c + \varepsilon_a \\ Y_b = -Y_c + \varepsilon_b \end{cases}$$

Conditional on Y_c, the expression level of gene C, the expression levels of genes A and B are independent.

(marginal) dependence



conditional independence



(Conditional) independence

Equivalent definitions

The random variables $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are *conditional independent* on $\mathbf{Y}_c$ if and only if:

$$f_{\mathbf{Y}_a \mid \mathbf{Y}_b, \mathbf{Y}_c}(\mathbf{y}_a, \mathbf{y}_b, \mathbf{y}_c) \quad = \quad f_{\mathbf{Y}_a \mid \mathbf{Y}_c}(\mathbf{y}_a, \mathbf{y}_c)$$

Thus, the conditional independence of $\mathbf{Y}_a$ and $\mathbf{Y}_b$ implies that $\mathbf{Y}_b$ can be excluded from the conditioning set.

The random variables $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are *conditional independent* on $\mathbf{Y}_c$ if and only if:

$$\begin{aligned} & f_{\mathbf{Y}_a, \mathbf{Y}_b, \mathbf{Y}_c}(\mathbf{y}_a, \mathbf{y}_b, \mathbf{y}_c) \\ & \qquad = \quad f_{\mathbf{Y}_a, \mathbf{Y}_c}(\mathbf{y}_a, \mathbf{y}_c)\, f_{\mathbf{Y}_b, \mathbf{Y}_c}(\mathbf{y}_b, \mathbf{y}_c) / f_{\mathbf{Y}_c}(\mathbf{y}_c) \end{aligned}$$

CI can be expressed in terms of the marginal densities.

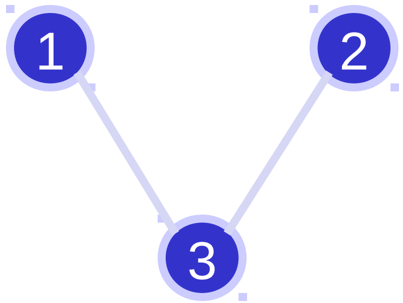
(Conditional) independence

Consider a pathway comprising of three genes.

Expr. levels of genes 1 and 2 are *independent conditional* on those of gene 3*:*

$$Y_1 \perp\!\!\!\perp Y_2 \,|\, Y_3$$

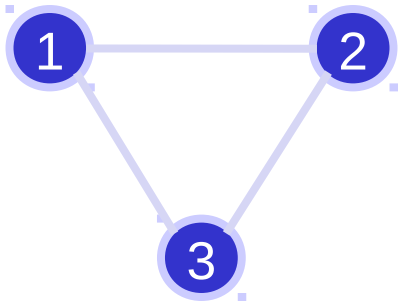
Graph:



Expr. levels of genes 1 and 2 are *dependent conditional* on those of gene 3*:*

$$Y_1 \not\!\perp\!\!\!\perp Y_2 \,|\, Y_3$$

Graph:

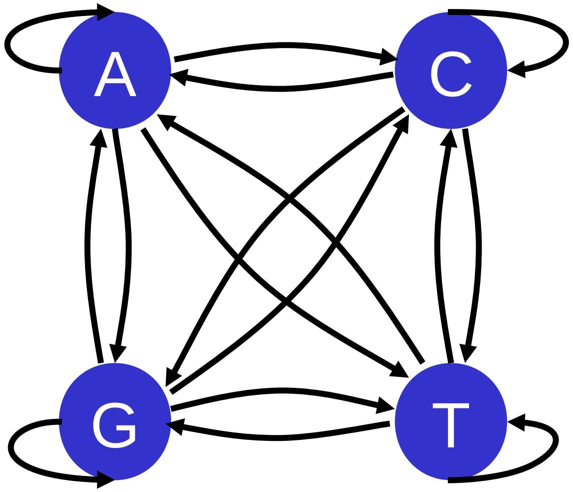


(Conditional) independence

Recall (from the first 4 lectures)

DNA sequence modeled by a 1st order Markov chain:

$$\mathbf{P} = \begin{array}{cc} & \overbrace{\begin{array}{cccc} A & C & G & T \end{array}}^{to} \\ \text{from} \left\{ \begin{array}{c} A \\ C \\ G \\ T \end{array} \right. & \begin{pmatrix} 0.1 & 0.1 & 0.1 & 0.7 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.4 & 0.3 & 0.2 & 0.1 \\ 0.1 & 0.3 & 0.5 & 0.1 \end{pmatrix} \end{array}$$



Question

Are X_t and X_{t+1} conditional independent on X_{t+2}?

What about X_t and X_{t+2} on X_{t+1}?

(Conditional) independence

Proposition (factorisation criterion)
Let $\mathbf{Y}_a$, $\mathbf{Y}_b$ and $\mathbf{Y}_c$ be *p-, q-,* and *r*-dimensional random variables. Then, $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are independent conditional on $\mathbf{Y}_c$, i.e.:

$$\mathbf{Y}_a \perp\!\!\!\perp \mathbf{Y}_b \,|\, \mathbf{Y}_c$$

if and only if there exists functions *g*(·) and *h*(·) such that:

$$f_{\mathbf{Y}_a, \mathbf{Y}_b, \mathbf{Y}_c}(\mathbf{y}_a, \mathbf{y}_b, \mathbf{y}_c) \quad = \quad g(\mathbf{y}_a, \mathbf{y}_c)\, h(\mathbf{y}_b, \mathbf{y}_c)$$

for all values of $\mathbf{y}_a$, $\mathbf{y}_b$ and all $\mathbf{y}_c$ with $f_{\mathbf{Y}_c}(\mathbf{y}_c) > 0$.

Note: functions *g*() and *f*() need not be densities.

(Conditional) independence

Once a conditional independence is known, others can be concluded to hold by application of the properties, e.g.:

i) $\mathbf{Y}_a \perp\!\!\!\perp \mathbf{Y}_b \,|\, \mathbf{Y}_c$ implies $\mathbf{Y}_b \perp\!\!\!\perp \mathbf{Y}_a \,|\, \mathbf{Y}_c$

ii) $\mathbf{Y}_a \perp\!\!\!\perp (\mathbf{Y}_b, \mathbf{Y}_c) \,|\, \mathbf{Y}_d$ implies $\mathbf{Y}_a \perp\!\!\!\perp \mathbf{Y}_b \,|\, \mathbf{Y}_d$

iii) $\mathbf{Y}_a \perp\!\!\!\perp (\mathbf{Y}_b, \mathbf{Y}_c) \,|\, \mathbf{Y}_d$ implies $\mathbf{Y}_a \perp\!\!\!\perp \mathbf{Y}_b \,|\, (\mathbf{Y}_c, \mathbf{Y}_d)$

where $\mathbf{Y}_a$,$\mathbf{Y}_b$,$\mathbf{Y}_c$, and $\mathbf{Y}_d$ random variables of arbitrary dimensions.

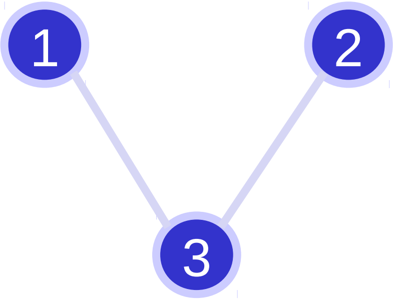
(Conditional) independence

Let $\mathbf{Y}$ be a *p*-dimensional random variable and $\mathcal{V} = \{1, \ldots, p\}$ the corresponding set of nodes.

The *conditional independence graph* of $\mathbf{Y}$ is an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ s.t.

$$(j_1, j_2) \notin \mathcal{E} \iff Y_{j_1} \perp\!\!\!\perp Y_{j_2} \,|\, Y_{\mathcal{V} \setminus \{j_1, j_2\}}$$

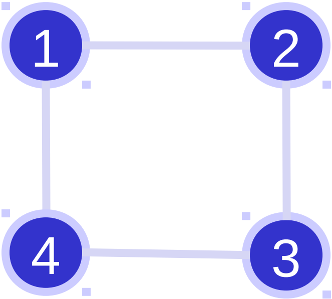
Example



$$\Longleftrightarrow \quad Y_1 \perp\!\!\!\perp Y_2 \,|\, Y_3$$

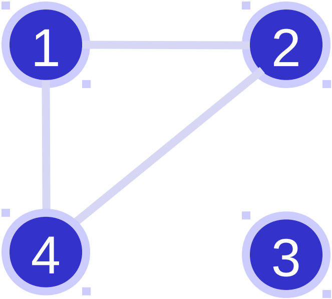
(Conditional) independence

Examples



⇔ e.g.

$$Y_1 \perp\!\!\!\perp Y_3 \mid \{Y_2, Y_4\}$$
$$Y_2 \perp\!\!\!\perp Y_4 \mid \{Y_1, Y_3\}$$
$$Y_1 \not\perp\!\!\!\perp Y_2 \mid \{Y_3, Y_4\}$$



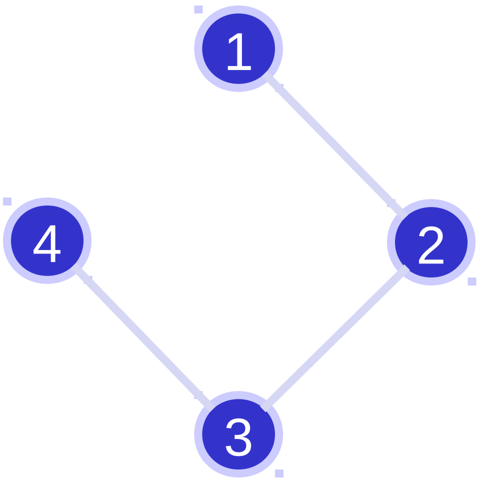
⇔ e.g.

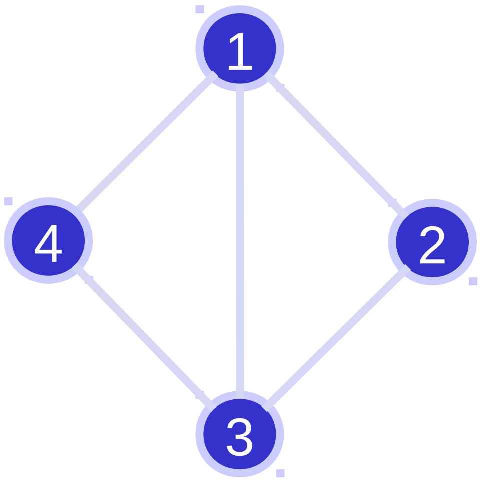
$$Y_1 \perp\!\!\!\perp Y_3 \mid \{Y_2, Y_4\}$$
$$Y_1 \perp\!\!\!\perp Y_3$$
$$Y_2 \not\perp\!\!\!\perp Y_4 \mid \{Y_1, Y_3\}$$

(Conditional) independence

Question

Which conditional independence relationships do the conditional independence graphs below convey?

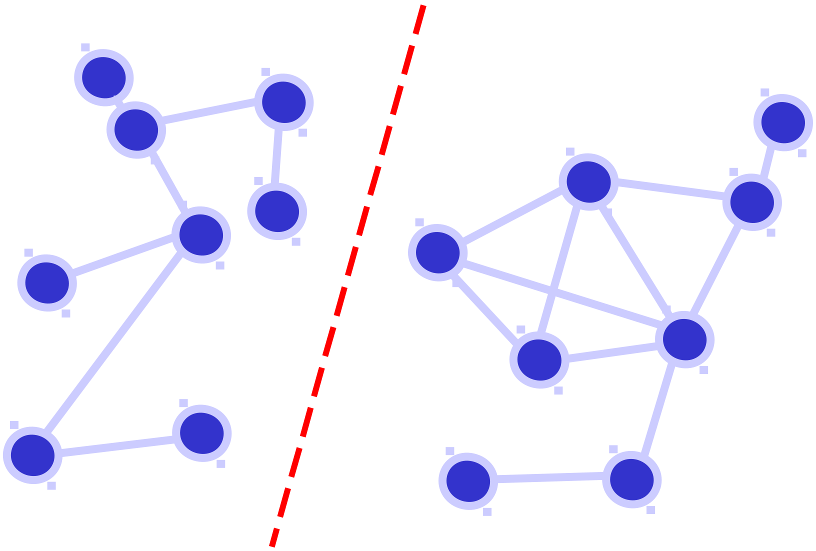




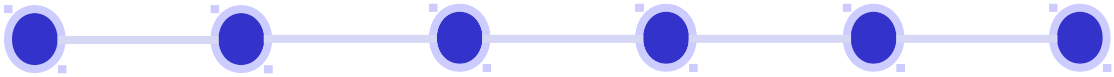
(Conditional) independence

Relevance

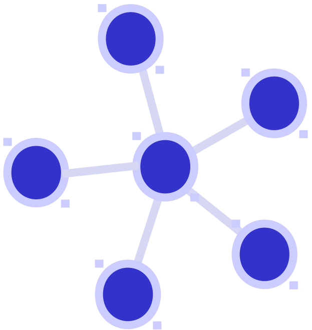
The pathway actually comprises two “sub-pathways”:



Chain graph (“signal processing”):



Star graph (“hub gene”):

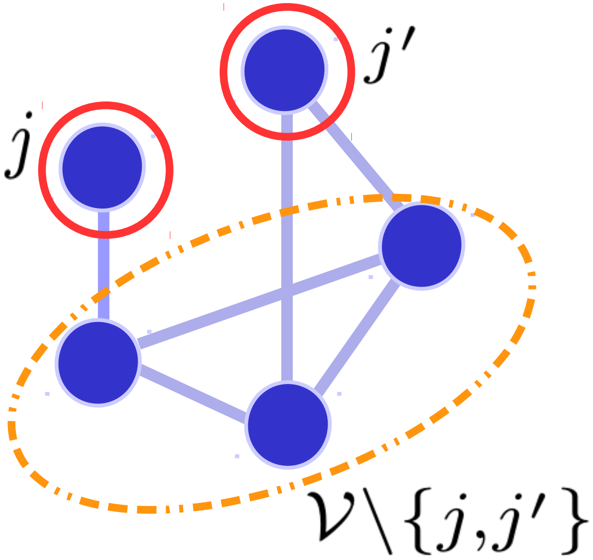


(Conditional) independence

Pairwise Markov property

Random variables of non-adjacent nodes j and j' are conditionally independent given the remaining random variables:

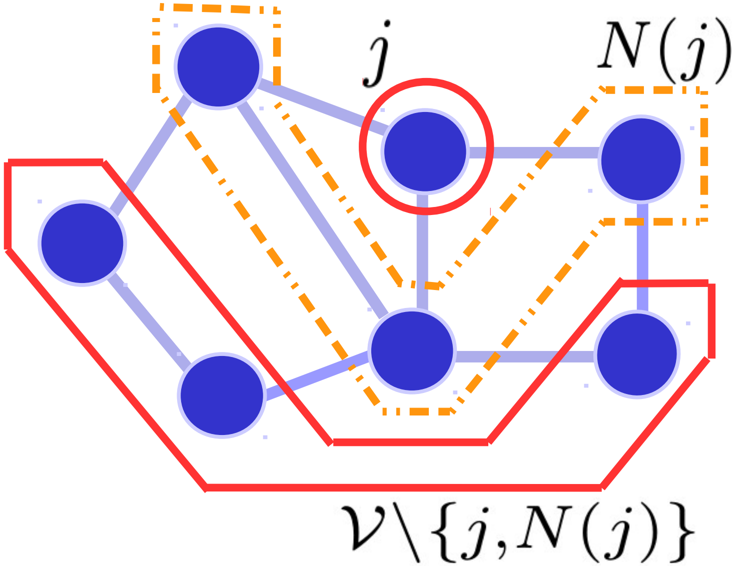
$$Y_j \perp\!\!\!\perp Y_{j'} | \mathbf{Y}_{\mathcal{V}\backslash\{j,j'\}}$$



Local Markov property

A random variable of node j is conditionally independent of all other random variables given those of its neighboring nodes *N*(j):

$$Y_j \perp\!\!\!\perp \mathbf{Y}_{N(j)} | \mathbf{Y}_{\mathcal{V}\backslash\{j,N(j)\}}$$

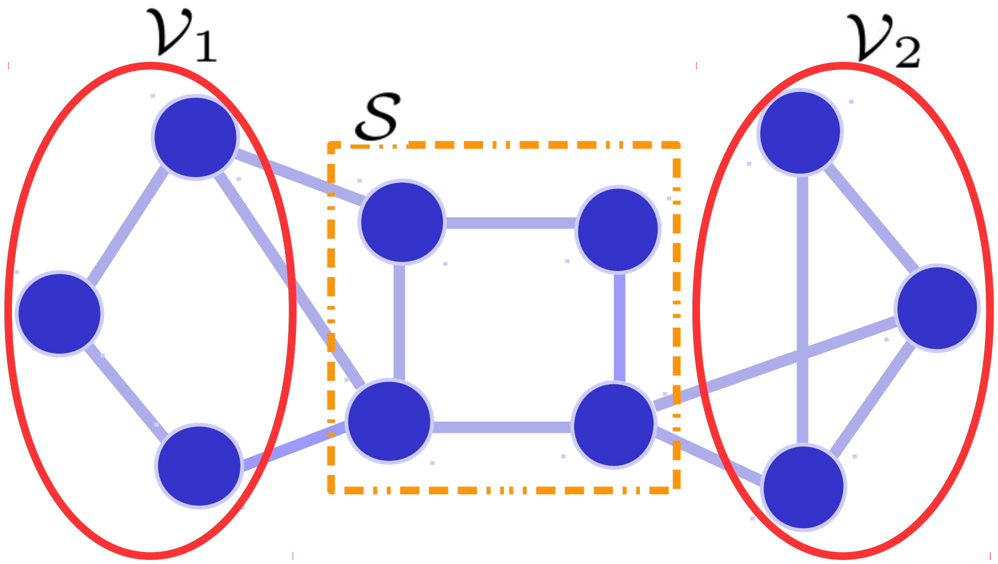


(Conditional) independence

Global Markov property

Two mutually exclusive subsets of random variables are conditionally independent given those of a separating subset:

$$\mathbf{Y}_{\mathcal{V}_1} \perp\!\!\!\perp \mathbf{Y}_{\mathcal{V}_2} | \mathbf{Y}_{\mathcal{S}}$$



Separating subset S:
All paths in graph *G* between node sets V_1 and V_2 run through *S*.

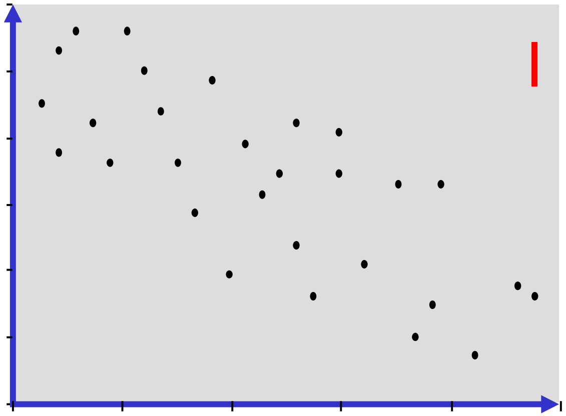
Theorem

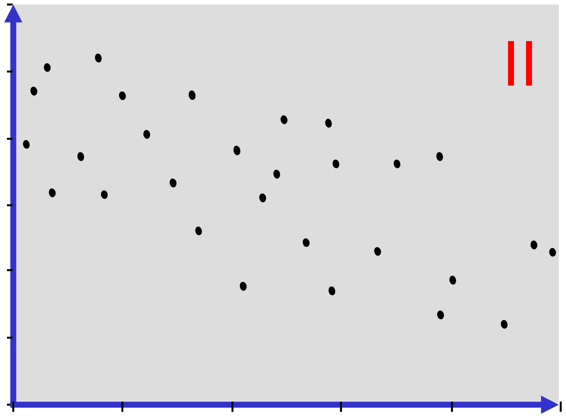
Under assumptions (that apply here) the pairwise, local and global Markov properties are equivalent.

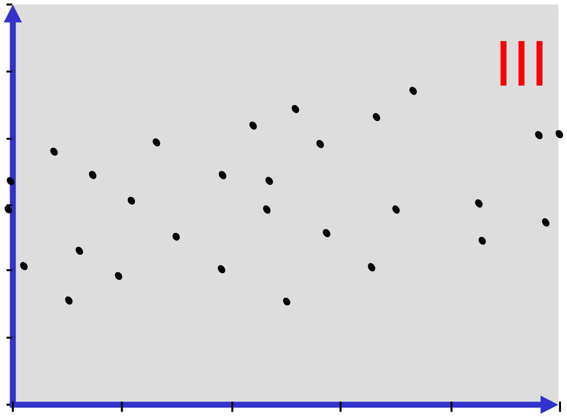
Covariance and correlation

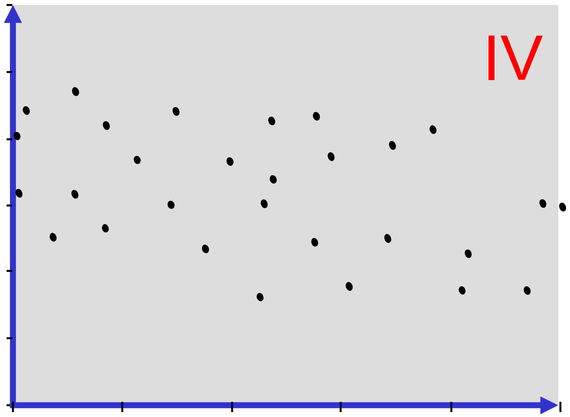
Covariance and correlation

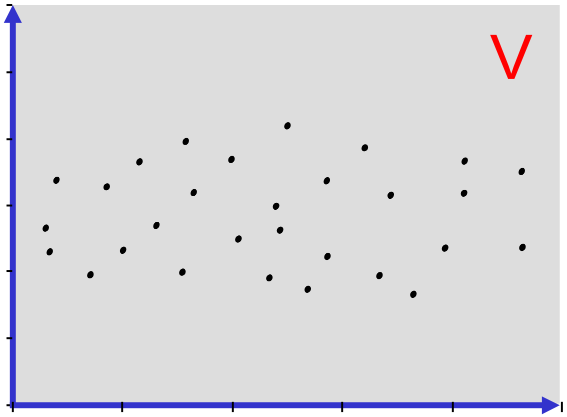
Scatterplots of data on two random variables.
Which show association?

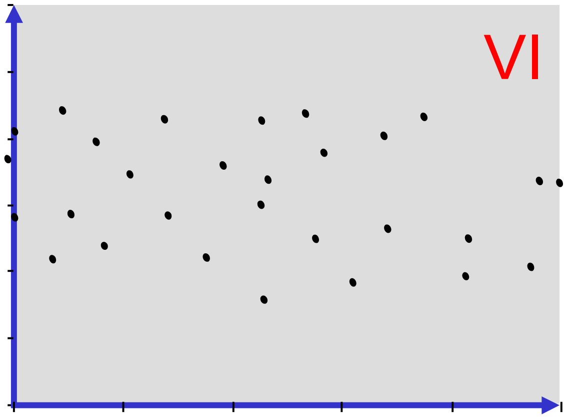






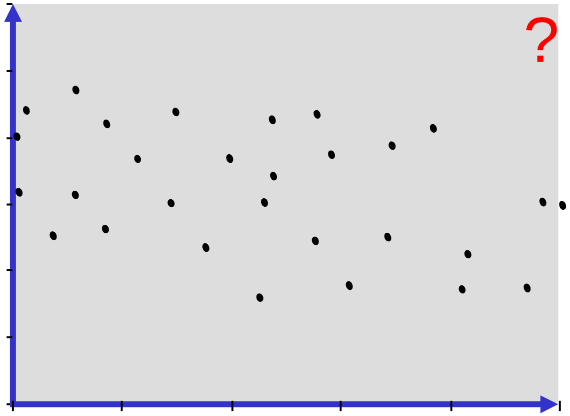






Covariance and correlation

Association between two random variables may be assessed graphically. This is not very exact and in boundary cases difficult to reach consensus.



Ideally, a measure of interrelatedness of the two variables.

Covariance is such a measure. It measures whether a positive deviation from the mean in one variable systemically coincides with a positive (or negative) deviation from the mean in another variable.

Covariance and correlation

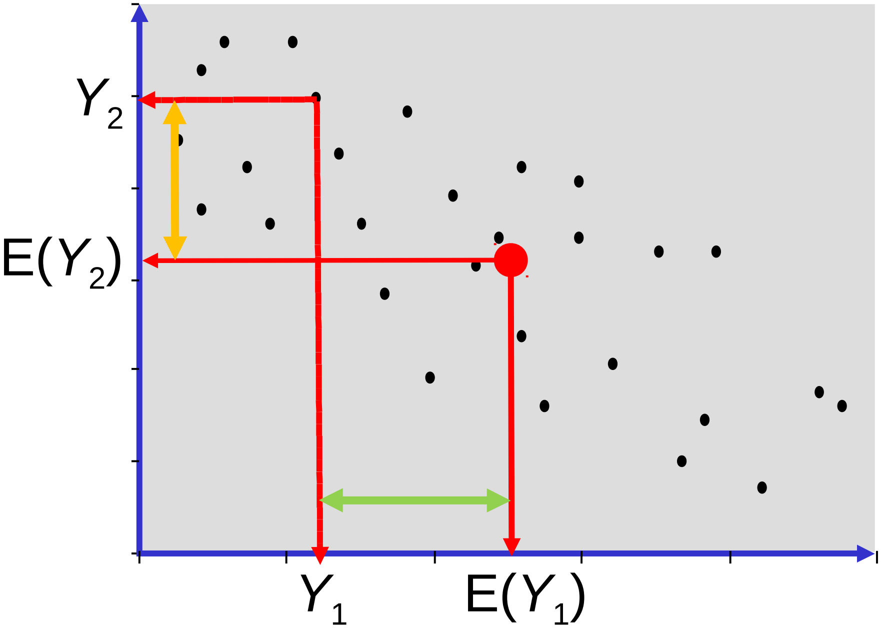
Covariance measures the linear dependence between two random variables.

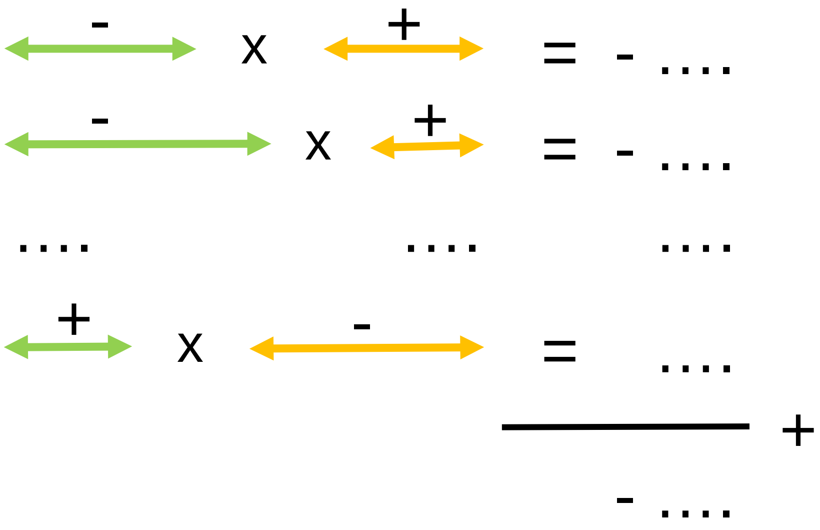
The covariance between random variables Y_1 and Y_2 is:

$$\mathrm{Cov}(Y_1, Y_2) = E\{\underbrace{[Y_1 - E(Y_1)]}\,\underbrace{[Y_2 - E(Y_2)]}\}$$

deviation from mean

estimation



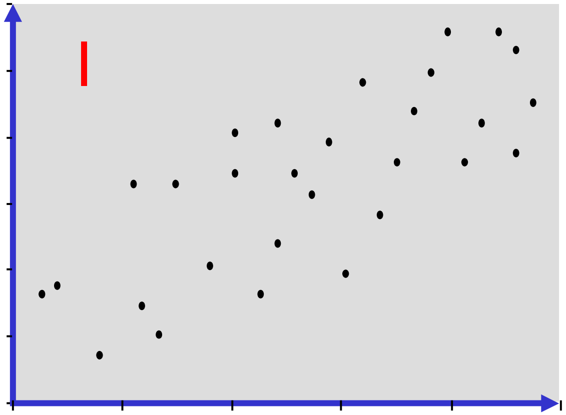


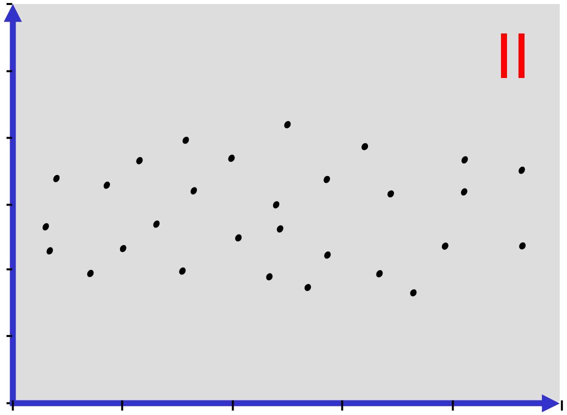
Covariance and correlation

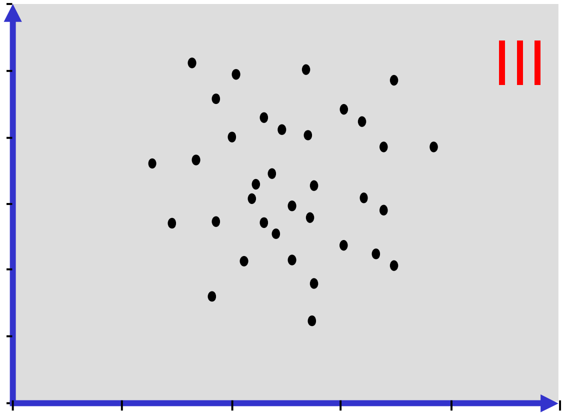
Question

Consider the expression levels of two genes.

What will be the estimated covariance between Y_1 and Y_2?







Covariance and correlation

Covariance properties (I)

Let Y_1 and Y_2 be two independent random variables and c a constant. Then:

$$\begin{aligned} \mathrm{Cov}(c, Y_2) &= 0 \\ \mathrm{Cov}(Y_1, Y_2) &= 0 \end{aligned}$$

Let Y_1 and Y_2 be two random variables. Then:

$$\begin{aligned} \mathrm{Cov}(Y_1, Y_2) &= \mathrm{Cov}(Y_2, Y_1) \\ \mathrm{Cov}(Y_1, Y_1) &= \mathrm{Var}(Y_1) \end{aligned}$$

Question: proof! (Hint: use definition of covariance).

Covariance and correlation

Covariance properties (II)

Let Y_1, Y_2, Y_3, and Y_4, be two random variables and *a* and *b* constants. Then:

$$\begin{aligned} \mathrm{Cov}(aY_1, bY_2) &= ab\mathrm{Cov}(Y_1, Y_2) \\ \mathrm{Cov}(Y_1 + a, Y_2 + b) &= \mathrm{Cov}(Y_1, Y_2) \end{aligned}$$

and

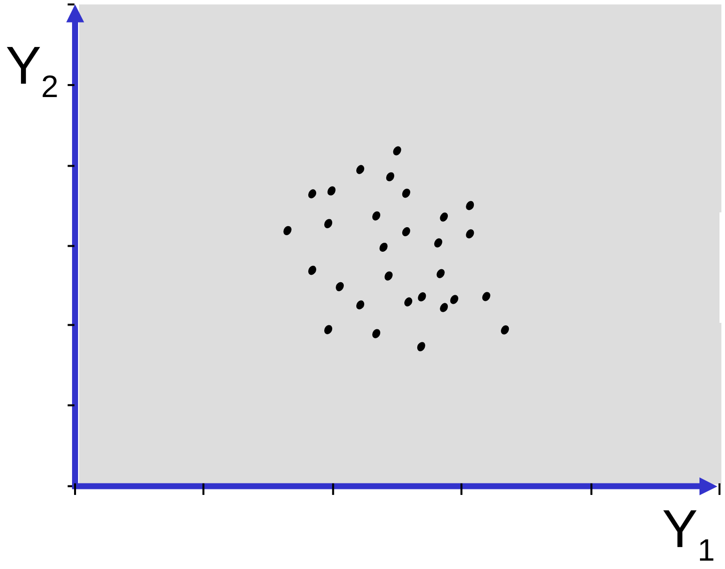
$$\begin{aligned} &\mathrm{Cov}(Y_1 + Y_2, Y_3 + Y_4) \\ &\quad = \mathrm{Cov}(Y_1, Y_3) + \mathrm{Cov}(Y_1, Y_4) \\ &\qquad + \mathrm{Cov}(Y_2, Y_3) + \mathrm{Cov}(Y_2, Y_4) \end{aligned}$$

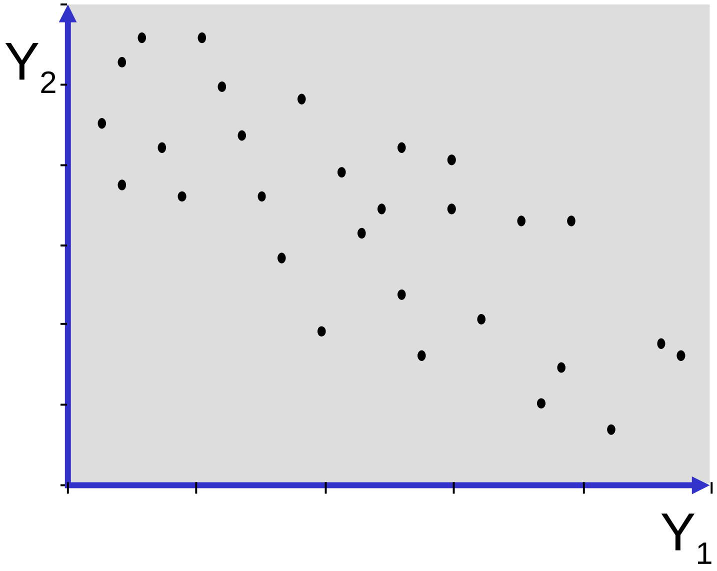
Covariance and correlation

Example

$$\begin{cases} \varepsilon_1, \varepsilon_2 & \sim & \mathcal{N}(0, \sigma_\varepsilon^2) \text{ i.i.d.} \\ Y_1 & = & \varepsilon_1 \\ Y_2 & = & \varepsilon_2 \end{cases}$$

$$\begin{cases} \varepsilon_1, \varepsilon_2 & \sim & \mathcal{N}(0, \sigma_\varepsilon^2) \text{ i.i.d.} \\ Y_1 & = & \varepsilon_1 \\ Y_2 & = & \beta Y_1 + \varepsilon_2 \end{cases}$$





$$\mathrm{Cov}(Y_1, Y_2) = \mathrm{Cov}(\varepsilon_1, \varepsilon_2) = 0$$

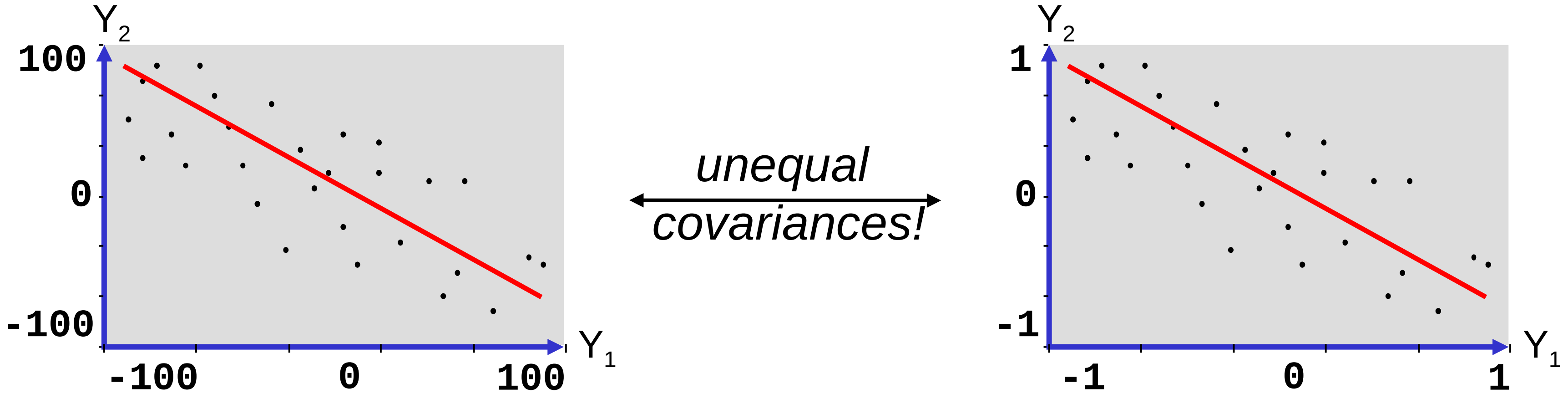
$$\mathrm{Cov}(Y_1, Y_2) = \text{???}$$

Covariance and correlation

Example

$$\left\{\begin{array}{rcl} \varepsilon_1, \varepsilon_2 & \sim & \mathcal{N}(0, \sigma_\varepsilon^2) \text{ i.i.d.} \\ Y_1 & = & \varepsilon_1 \\ Y_2 & = & \beta Y_1 + \varepsilon_2 \end{array}\right. \qquad \longrightarrow \qquad \operatorname{Cov}(Y_1, Y_2) = \beta\sigma_\varepsilon^2$$

Covariance thus depends on variance of Y_1, while linear relation (β) between Y_1 and Y_2 is unchanged.



unequal covariances!

Undesirable property for a measure of linear dependence.

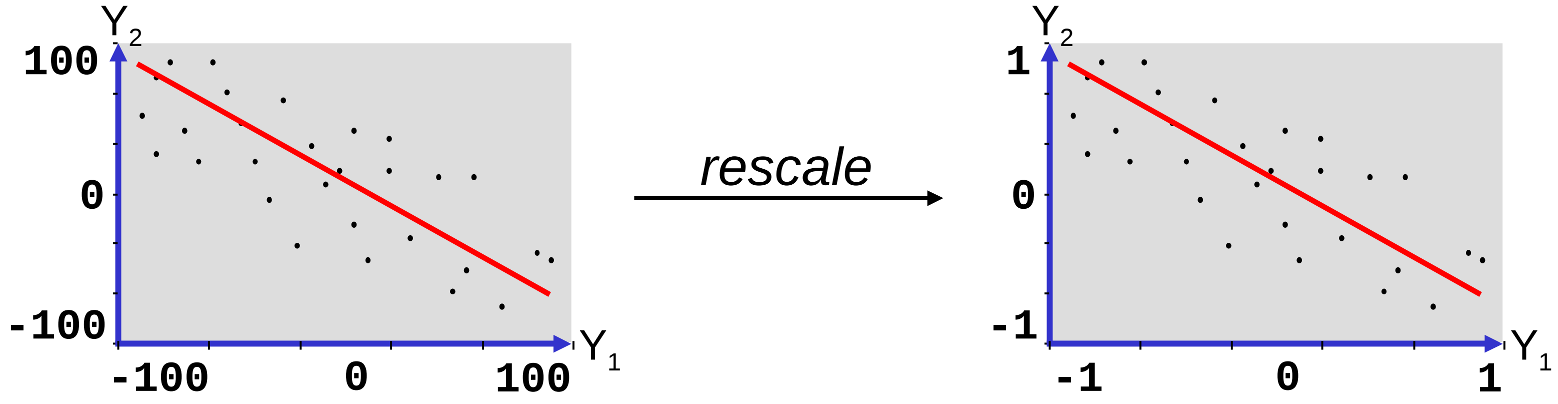
Covariance and correlation

Solution

Standardization of Y_1 and Y_2

$$\text{Cov}(\tilde{Y}_1, \tilde{Y}_2) = \text{Cov}(a_1 Y_1, a_2 Y_2) = a_1 a_2 \text{Cov}(Y_1, Y_2)$$

with $a_j = [\text{Var}(Y_j)]^{-1/2}$



$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= \mathbf{8483.662} \\ \text{Cov}(\tilde{Y}_1, \tilde{Y}_2) &= \mathbf{0.675} \end{aligned}$$

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= \mathbf{1.007} \\ \text{Cov}(\tilde{Y}_1, \tilde{Y}_2) &= \mathbf{0.697} \end{aligned}$$

Covariance and correlation

Pearson's correlation coefficient

Normalized covariance between Y_1 andY_2:

$$\rho(Y_1, Y_2) = \mathrm{Cor}(Y_1, Y_2) = \frac{\mathrm{Cov}(Y_1, Y_2)}{\sqrt{\mathrm{Var}(Y_1)}\sqrt{\mathrm{Var}(Y_2)}}$$

It measures the degree of linear dependence between the two random variables Y_1 and Y_2.

$\rho(Y_1, Y_2)$ in [-1, 1], with

→ $\rho = 1$: perfect positive linear relationship.

→ $\rho = 0$: absence of linear dependency.

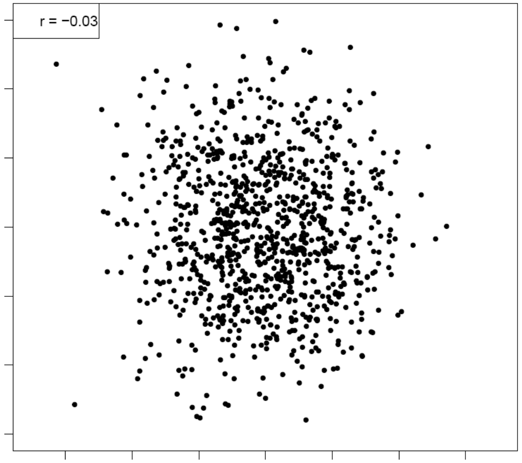
→ $\rho = -1$: perfect negative linear relationship.

Closer $|\rho|$ to one: stronger linear dependency.

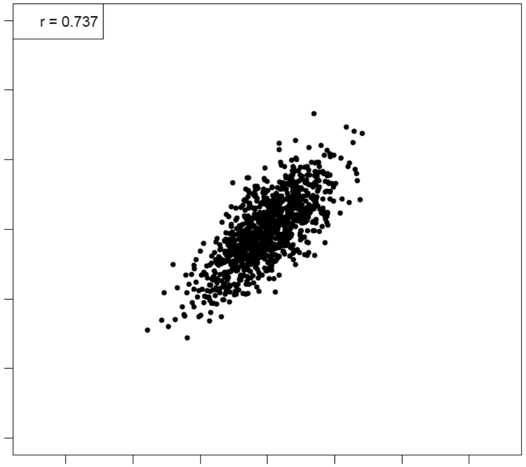
Covariance and correlation

Pearson’s correlation measures only linear dependence.

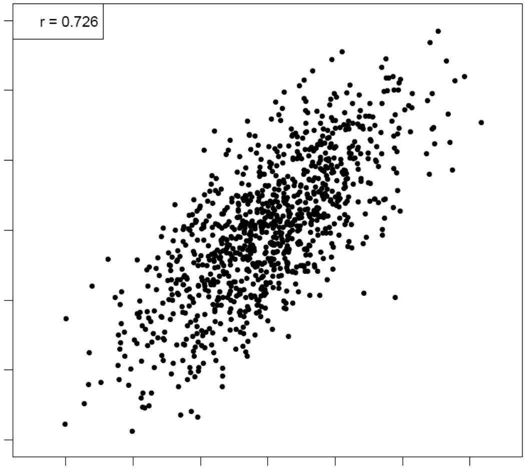
$r = -0.030$



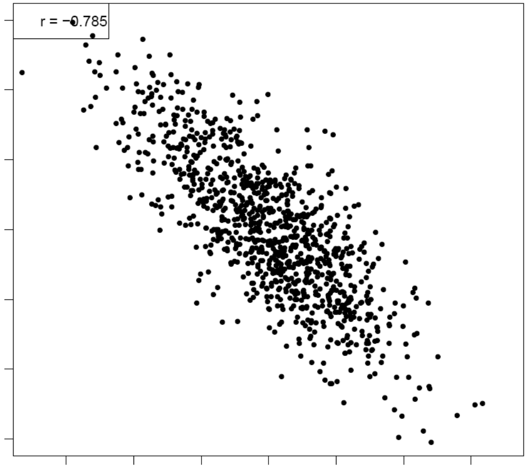
$r = 0.737$



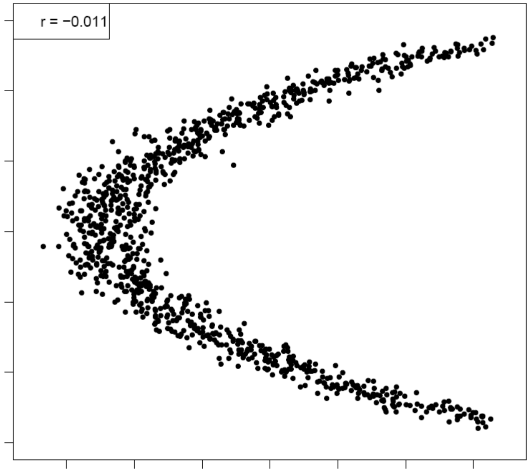
$r = 0.726$



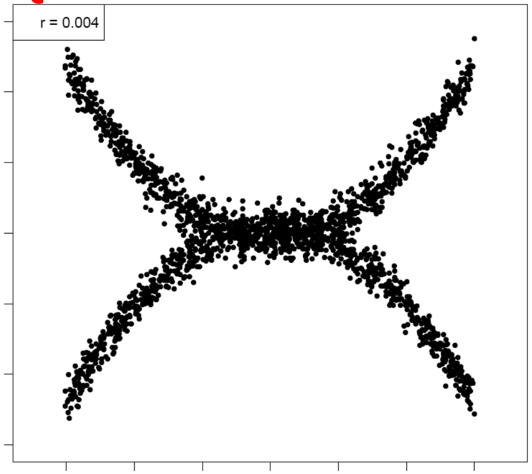
$r = -0.785$



$r = -0.011$



Question: $r \approx$??



Covariance and correlation

Clearly, a Pearson correlation coefficient equal to zero does not imply the absence of nonlinear relationships.

Question

Let $X \sim N(0, 1)$. Define Y through:

$Y = X$ if $|X| > c$
$Y = -X$ if $|X| < c$

with $c \geq 0$.

What is $\text{Cor}(X, Y)$ for $c=0.01$? For $c=1.5$? For $c=10$?

Covariance and correlation

Estimation

Pearson’s correlation coefficient is estimated by:

$$\hat{\rho}(Y_1, Y_2) \quad = \quad \frac{\widehat{\mathrm{Cov}}(Y_1, Y_2)}{\sqrt{\widehat{\mathrm{Var}}(Y_1)}\sqrt{\widehat{\mathrm{Var}}(Y_2)}}$$

where

$$\widehat{\mathrm{Cov}}(Y_1, Y_2) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(Y_{i1} - \hat{\mu}_1)(Y_{i2} - \hat{\mu}_2)$$

$$\widehat{\mathrm{Var}}(Y_j) = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}(Y_{ij} - \hat{\mu}_j)^2$$

$$\hat{\mu}_j = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} Y_{ij}$$

Denoted r and called the *sample correlation coefficient*.

Covariance and correlation

Distribution

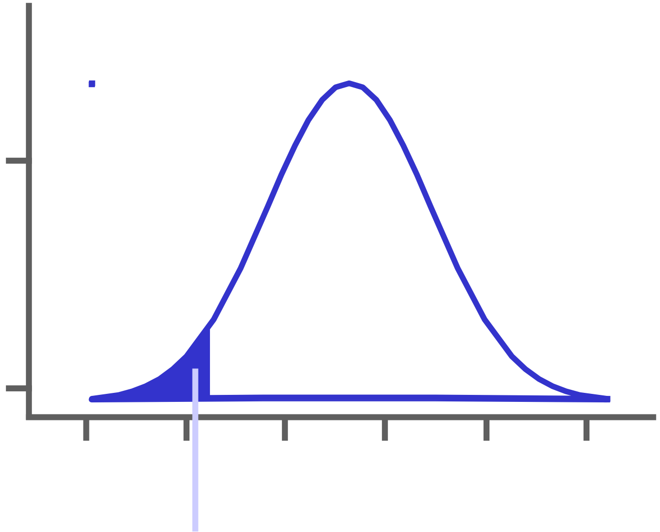
Under the assumption of a multivariate normal distribution, the Fisher transformed sample correlation coefficient:

$$F(\hat{\rho}) \;=\; \frac{1}{2}\log[(1+\hat{\rho})/(1-\hat{\rho})] \;=\; \operatorname{arctanh}(\hat{\rho})$$

follows approximately a normal distribution:

$$F(\hat{\rho}) \;\sim\; \mathcal{N}[F(\rho),(n-3)^{-1}]$$

Can now to test H_0: $\rho = 0$.



p-value = P (X ≤ x)

Covariance and correlation

Covariance matrix

The definition of covariance extends to random vectors:

$$
\begin{aligned}
&\mathrm{Cov}(\mathbf{X},\mathbf{Y})\\
&\quad = \quad E\{[\mathbf{X}-E(\mathbf{X})][\mathbf{Y}-E(\mathbf{Y})]^\top\}\\
&\quad = \quad \int_{\mathbb{R}^p}\int_{\mathbb{R}^p}[\mathbf{x}-E(\mathbf{X})][\mathbf{y}-E(\mathbf{Y})]^\top f_{(\mathbf{X},\mathbf{Y})}(\mathbf{x},\mathbf{y})\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{y}
\end{aligned}
$$

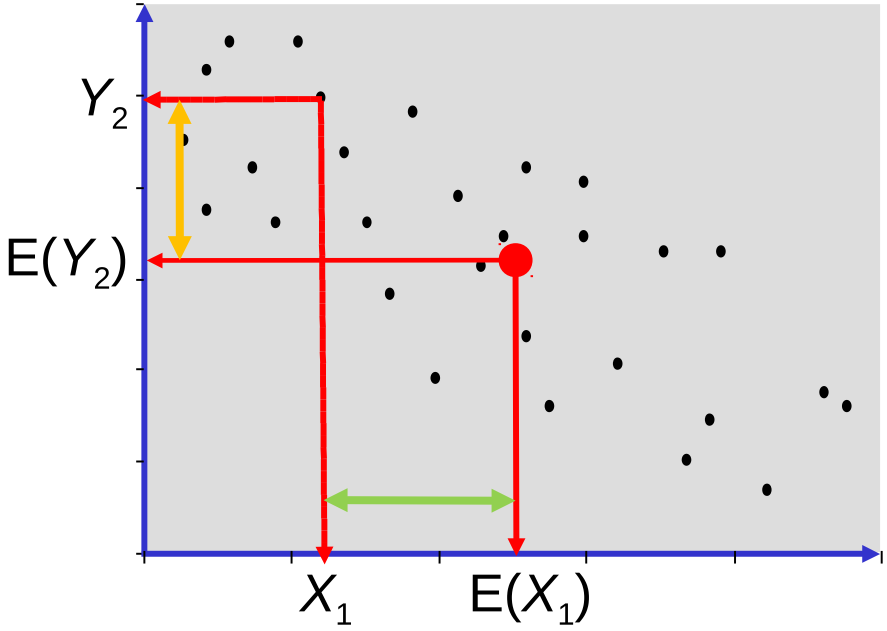
No longer a scalar, covariance is now a *pxp* matrix:

$$
\begin{pmatrix}
\mathrm{Cov}[(\mathbf{X})_1,(\mathbf{Y})_1] & \ldots & \mathrm{Cov}[(\mathbf{X})_1,(\mathbf{Y})_p]\\
\vdots & \ddots & \vdots\\
\mathrm{Cov}[(\mathbf{X})_p,(\mathbf{Y})_1] & \ldots & \mathrm{Cov}[(\mathbf{X})_p,(\mathbf{Y})_p]
\end{pmatrix}
$$

Covariance and correlation

Covariance matrix

The elements of a covariance matrix are the pairwise covariances of the elements of random vectors **X** and **Y**:



$$
\begin{aligned}
[\mathrm{Cov}(\mathbf{X},\mathbf{Y})]_{1,2} & \\
&= \mathrm{Cov}[(\mathbf{X})_1,(\mathbf{Y})_2] \\
&= \mathrm{Cov}(X_1,Y_2) \\
&= E\{[X_1 - E(X_1)] \\
&\qquad [Y_2 - E(Y_2)]\}
\end{aligned}
$$

Covariance and correlation

Question

Consider the random vector $\mathbf{Y} = (Y_1, Y_2, Y_3)^\top$ with covariance matrix:

$$\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}) = \begin{pmatrix} 2 & 0 & -1 \\ 0 & 3 & 1 \\ -1 & 1 & 1 \end{pmatrix}$$

→ What is the meaning of the diagonal elements?
→ Why is the above matrix symmetric?
→ What does the value of (1,2) element imply?

Covariance and correlation

Covariance matrix properties (I)

Let **X** and **Y** be two independent multivariate random variables. Then:

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = 0$$

Let **X** be a multivariate random variable and **c** a vector with constants. Then:

$$\mathrm{Cov}(\mathbf{c}, \mathbf{X}) = 0$$

Let **Y** be a multivariate random variable. Then:

$$\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}) = \mathrm{Var}(\mathbf{Y})$$

Covariance and correlation

Covariance matrix properties (II)

Let **W**, **X, Y** and **Z** be multivariate random variables. Then:

$$\begin{aligned}&\mathrm{Cov}(\mathbf{W}+\mathbf{X},\mathbf{Y}+\mathbf{Z})\\ &\quad= \mathrm{Cov}(\mathbf{W},\mathbf{Y})+\mathrm{Cov}(\mathbf{W},\mathbf{Z})\\ &\qquad+\mathrm{Cov}(\mathbf{X},\mathbf{Y})+\mathrm{Cov}(\mathbf{X},\mathbf{Z})\end{aligned}$$

Let **X** and **Y** be two multivariate random variables and **A** and **B** coefficient matrices. Then:

$$\mathrm{Cov}(\mathbf{A}\mathbf{X},\mathbf{B}\mathbf{Y}) = \mathbf{A}\,\mathrm{Cov}(\mathbf{X},\mathbf{Y})\mathbf{B}^{\mathrm{T}}$$

Covariance and correlation

Correlation matrix

Similarly, the correlation between two random vectors is:

$$\mathrm{Cor}(\mathbf{X},\mathbf{Y}) = \begin{pmatrix} \mathrm{Cor}[(\mathbf{X})_1,(\mathbf{Y})_1] & \dots & \mathrm{Cor}[(\mathbf{X})_1,(\mathbf{Y})_p] \\ \vdots & \ddots & \vdots \\ \mathrm{Cor}[(\mathbf{X})_p,(\mathbf{Y})_1] & \dots & \mathrm{Cor}[(\mathbf{X})_p,(\mathbf{Y})_p] \end{pmatrix}$$

with e.g.:

$$\mathrm{Cor}[(\mathbf{X})_1,(\mathbf{Y})_2] = \mathrm{Cor}(X_1,Y_2) = \frac{\mathrm{Cov}(X_1,Y_2)}{\sqrt{\mathrm{Var}(X_1)}\sqrt{\mathrm{Var}(Y_2)}}$$

The correlation matrix contains the pairwise correlations.

Covariance and correlation

Question

Consider the random vector $\mathbf{Y} = (Y_1, Y_2, Y_3)^\top$ with covariance matrix:

$$\mathrm{Cov}(\mathbf{Y}, \mathbf{Y}) = \begin{pmatrix} 2 & 0 & -1 \\ 0 & 3 & 1 \\ -1 & 1 & 1 \end{pmatrix}$$

Consider the related correlation matrix. What is on the diagonal of this correlation matrix?

Multivariate normal distribution

Multivariate normal distribution

Denote a *p*-dimensional $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)^{\mathrm{T}}$ random variable following a *multivariate normal distribution* by:

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

with a *mean* parameter:

$$\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^{\mathrm{T}} \in \mathbb{R}^p$$

and a *covariance* parameter $\boldsymbol{\Sigma} \in \mathcal{S}^p_{++}$:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix}$$

Multivariate normal distribution

Density

The *p*-variate normal distribution has density f(**Y**$_i$) equal to:

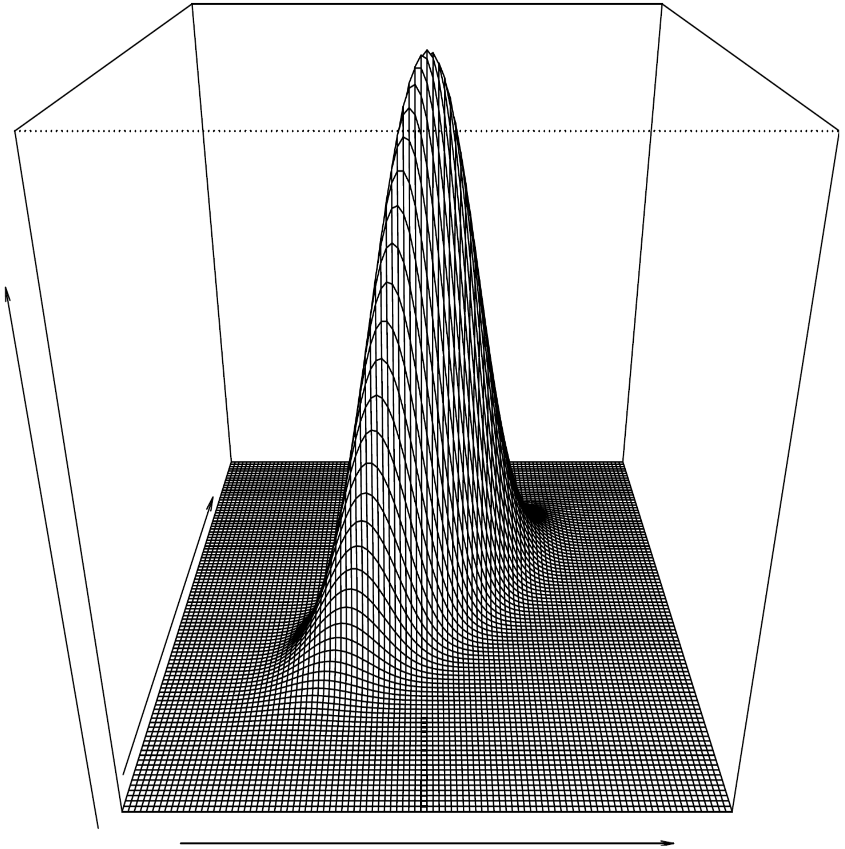
$$\frac{1}{(2\pi)^{p/2}\,|\boldsymbol{\Sigma}|^{1/2}}\,\exp\left[-\frac{1}{2}(\mathbf{Y}_i-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}\,(\mathbf{Y}_i-\boldsymbol{\mu})\right]$$

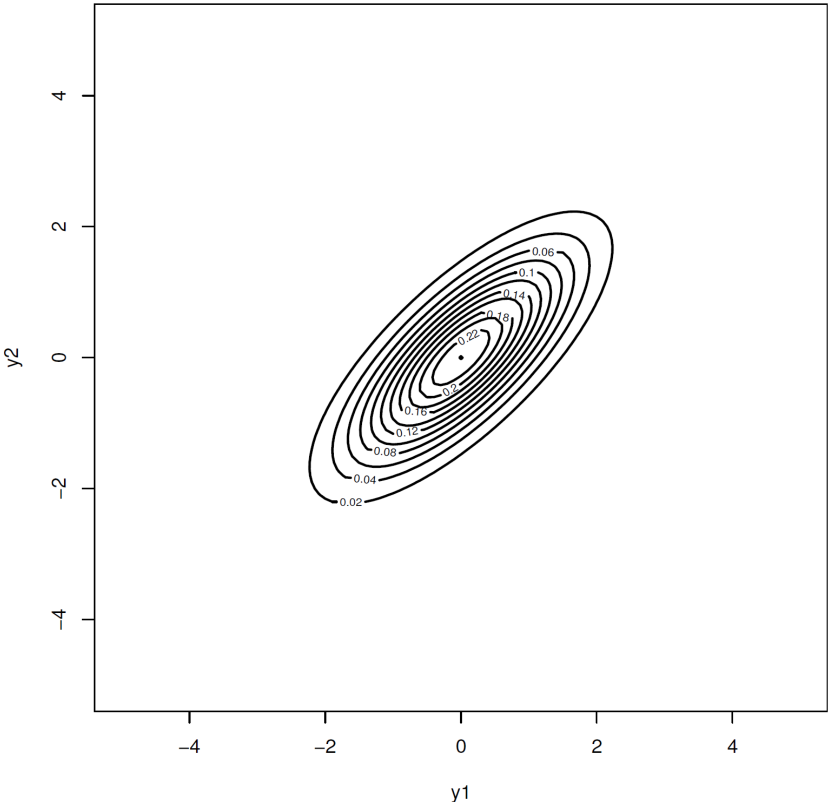
Recall the univariate normal distribution density:

$$\begin{aligned} f(Y_i) &= \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\left[-\frac{1}{2}(Y_i-\mu)^2/\sigma^2\right] \\ &= \frac{1}{(2\pi)^{1/2}\,\sigma}\,\exp\left[-\frac{1}{2}(Y_i-\mu)\,\sigma^{-2}\,(Y_i-\mu)\right] \end{aligned}$$

Multivariate normal distribution

The density of a bivariate (*p*=2) normal distribution.

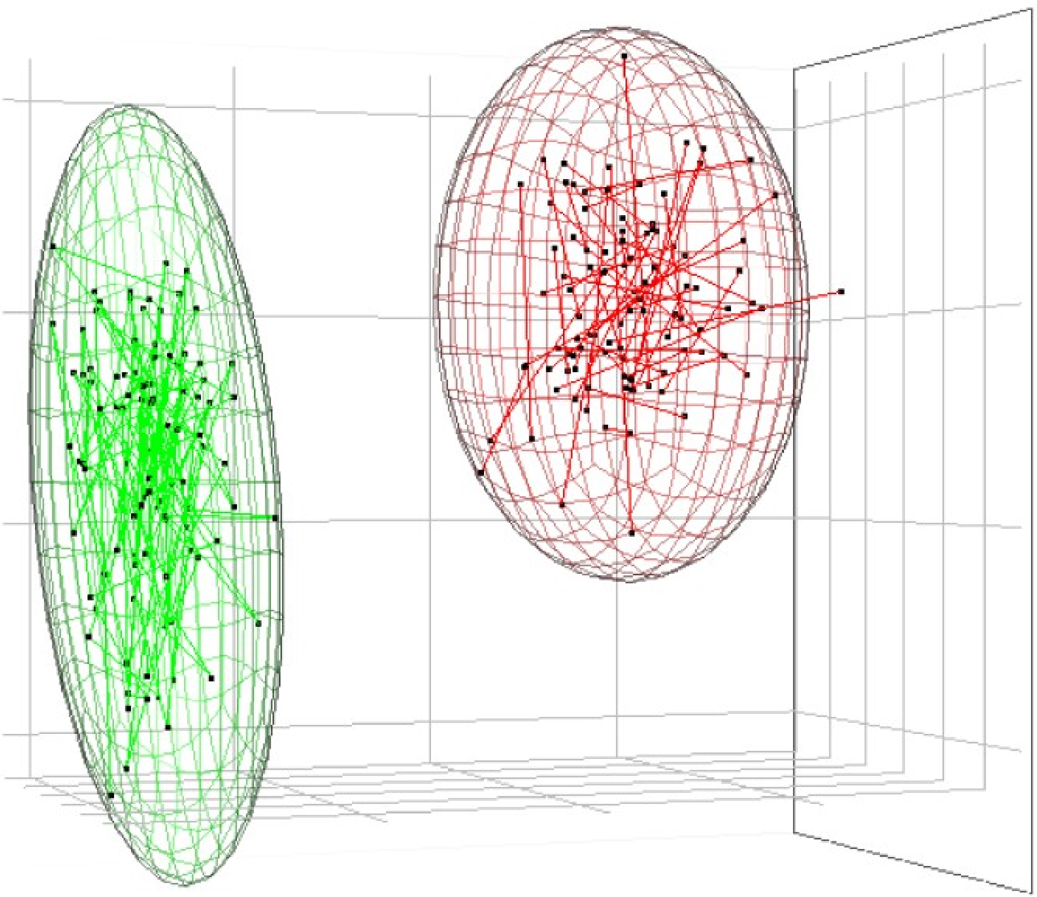


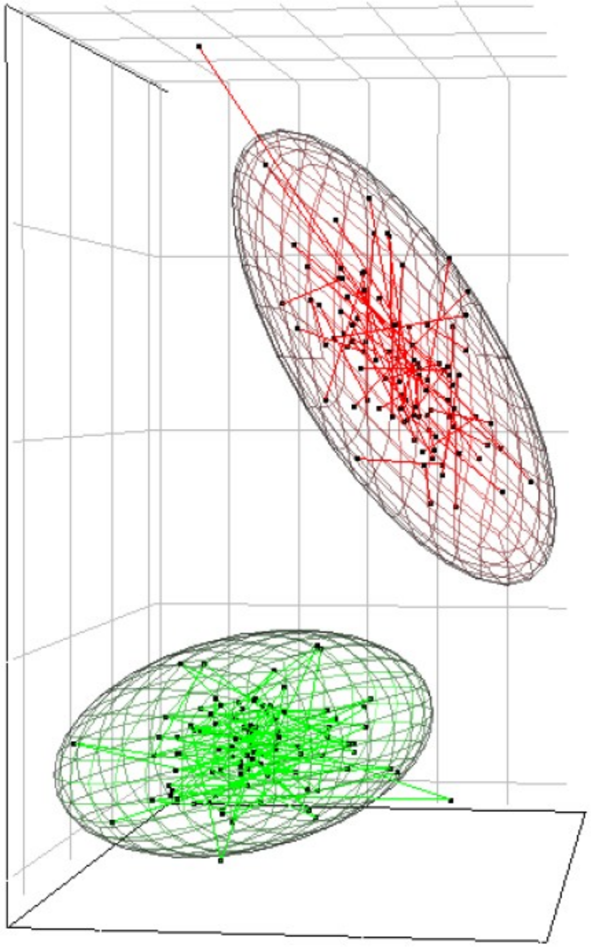


Density represented by level sets: {**Y** : f(**Y**) = c}. Observations with equal likelihood.

Multivariate normal distribution

Data distribution of trivariate (*p*=3) normal distributions.





Multivariate normal distribution

Standard multivariate normal

The random variable $\mathbf{Y} = (Y_1, Y_2, Y_3)^\top$ is standard normally distributed if:

$$\boldsymbol{\mu} = \mathbf{0}_{p\times 1} \quad \text{and} \quad \boldsymbol{\Sigma} = \mathbf{I}_{p\times p}$$

Thus:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix} \right)$$

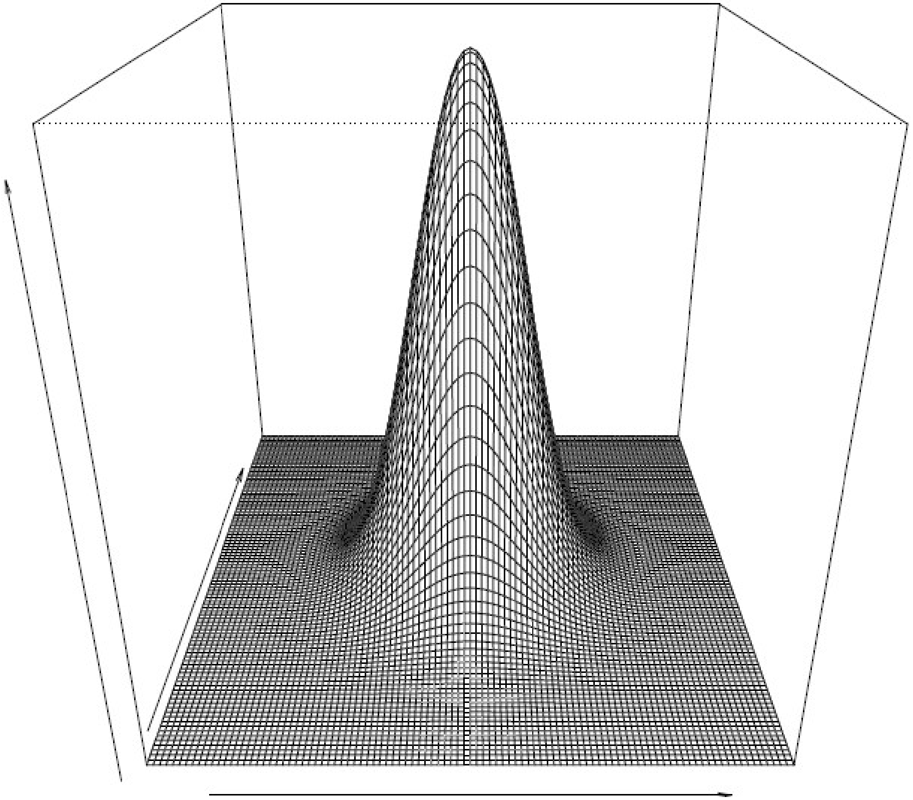
Put differently:

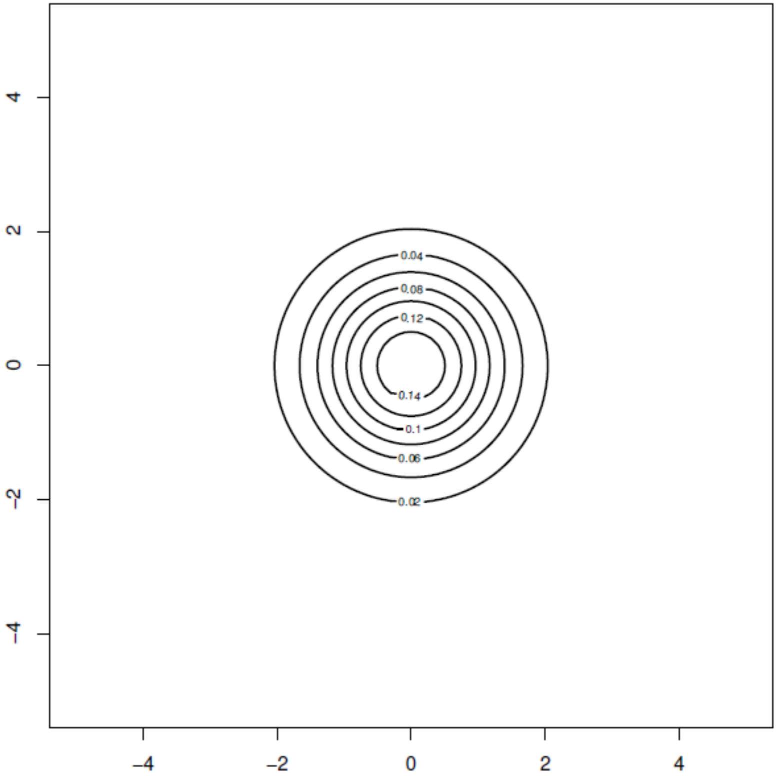
$$Y_j \text{ i.i.d. with } Y_j \sim \mathcal{N}(0,1)$$

Multivariate normal distribution

Standard bivariate normal

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \quad \leftrightarrow \quad \left\{ \begin{array}{c} Y_1 \sim \mathcal{N}(0,1), \\ Y_2 \sim \mathcal{N}(0,1), \\ Y_1 \perp\!\!\!\perp Y_2 \end{array} \right.$$





Multivariate normal distribution

Any multivariate normal random variable can be derived from the standard normal one.

Let $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \mathbf{I}_{p\times p})$, $\boldsymbol{\mu} \in \mathbb{R}^p$,
and $\mathbf{L} \in \mathcal{M}^p$ such that $\operatorname{rank}(\mathbf{L}) = p$

Now define:

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{L}\mathbf{Z}$$

i.e:

$$\begin{cases} Y_1 = \mu_1 + (\mathbf{L})_{11} Z_1 + (\mathbf{L})_{12} Z_2 + \ldots + (\mathbf{L})_{1p} Z_p \\ Y_2 = \mu_2 + (\mathbf{L})_{21} Z_1 + (\mathbf{L})_{22} Z_2 + \ldots + (\mathbf{L})_{2p} Z_p \\ \ldots \end{cases}$$

Then:

$$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{L}\mathbf{L}^\top)$$

Multivariate normal distribution

Question

Let the random variable $\mathbf{Y}$ be defined as on the previous slide. Verify:

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$$

and

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top \in \mathbf{S}_{++}^p ?$$

Hint (for part 2)

Use the singular value decomposition of $\mathbf{L}$:

$$\mathbf{L} = \mathbf{U}_\ell \mathbf{D}_\ell \mathbf{V}_\ell^\top$$

Multivariate normal distribution

Bivariate normal distribution.

Recall model:

$$\begin{cases} \varepsilon_1, \varepsilon_2 & \sim & \mathcal{N}(0, \sigma_\varepsilon^2) \text{ i.i.d.} \\ Y_1 & = & \varepsilon_1 \\ Y_2 & = & \beta Y_1 + \varepsilon_2 \end{cases}$$

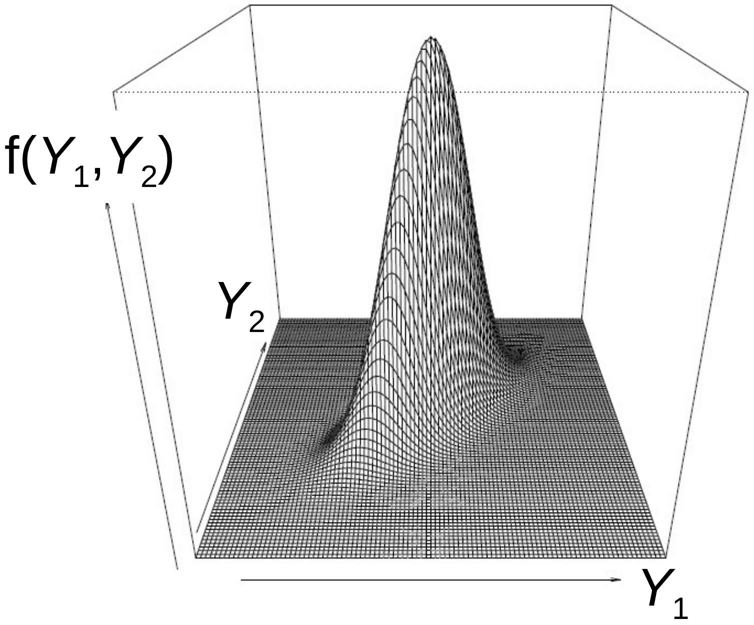
Then:

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

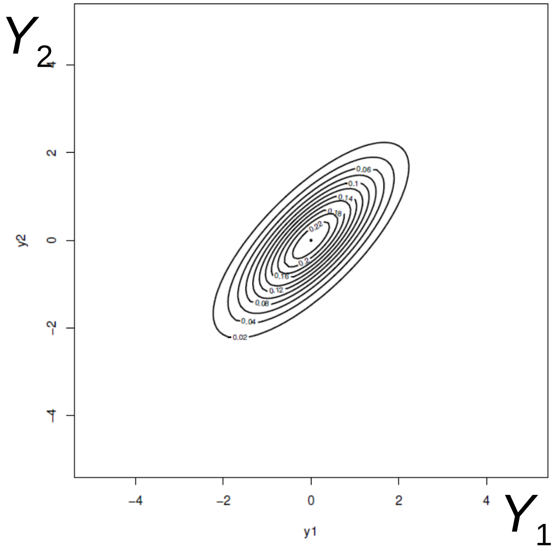
with:

$$\boldsymbol{\mu} = (0, 0)^\top$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_\varepsilon^2 & \beta\sigma_\varepsilon^2 \\ \beta\sigma_\varepsilon^2 & (1+\beta^2)\sigma_\varepsilon \end{pmatrix}$$



joint density



Multivariate normal distribution

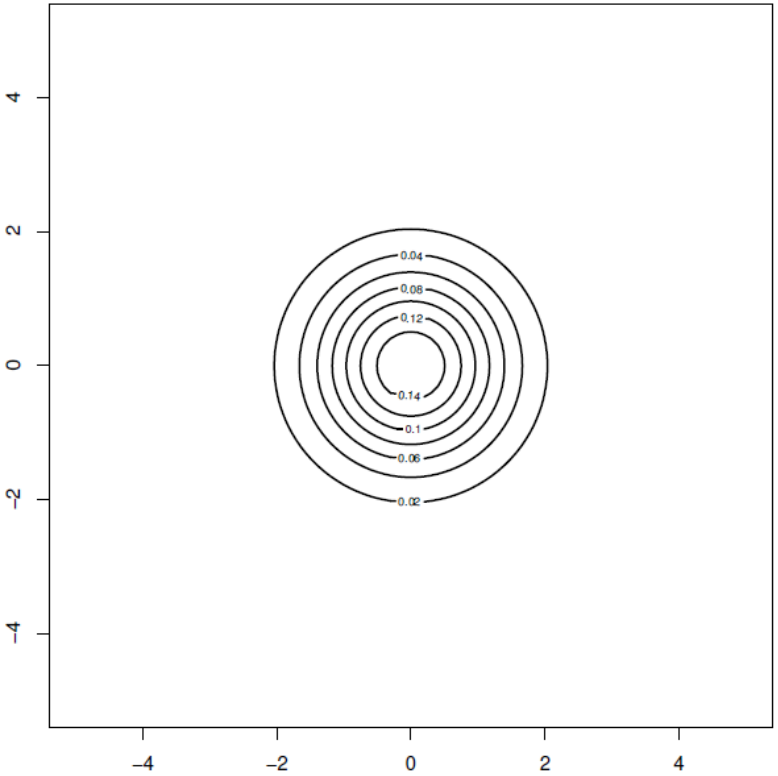
Question

Let Y be a bivariate normally distributed, random variable.

How would you calculate:

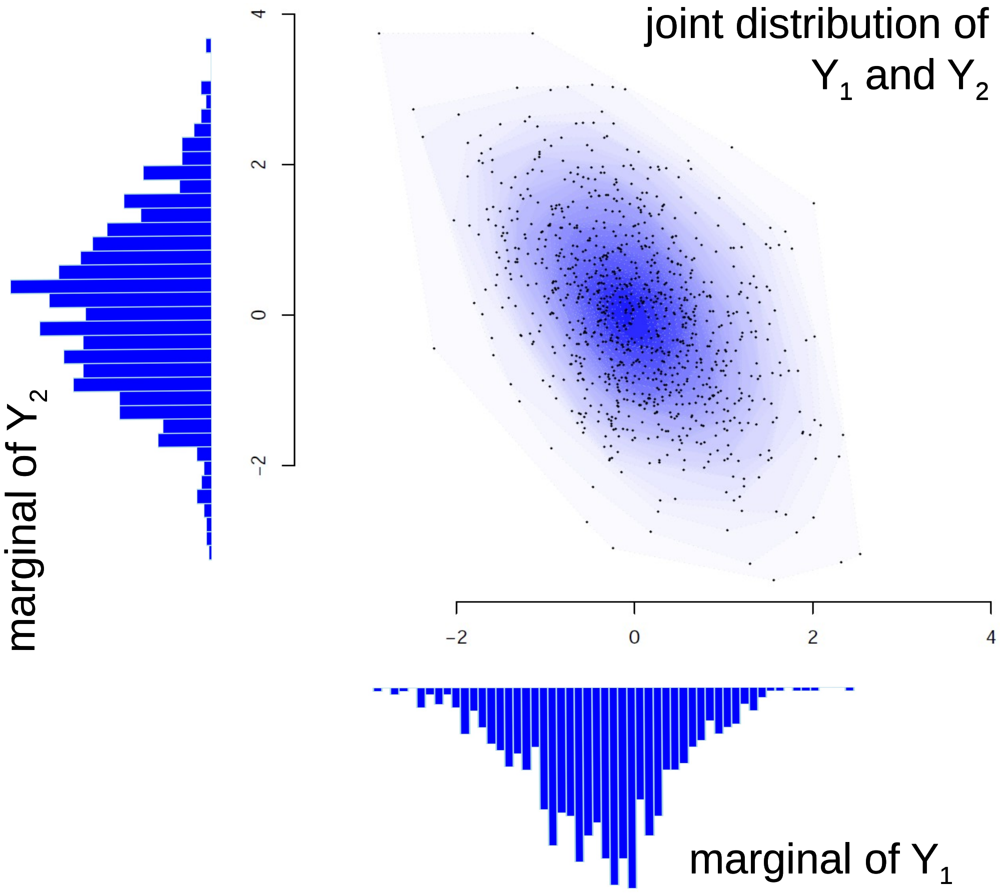
$P(Y_1 \geq 0)$

$P(Y_1 + Y_2 \geq 0)$



Multivariate normal distribution

The *marginal distribution* of a subset of random variables $Y_1, \ldots, Y_p$ is the distribution of random variables in the subset.



joint distribution of Y_1 and Y_2

marginal of Y_2

marginal of Y_1

Multivariate normal distribution

For the bivariate normal the marginal of Y_1 is:

$$f_{Y_1}(y_1) = \int_{\mathbb{R}} f_{(Y_1,Y_2)}(y_1, y_2) dy_2 = \frac{1}{\sqrt{2\pi}\sigma_1} \exp[-\frac{1}{2\sigma_1^2}(y_1 - \mu_1)^2]$$

Thus: $Y_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$, where, e.g.:

$$\mu_1 = \mathbb{E}(Y_1) = \int_{\mathbb{R}} y_1 f_{Y_1}(y_1) dy_1$$

This result (normality) also holds for $p > 2$.

Consequence

As the marginal distribution of a multivariate normal is itself (multivariate) normal, we can interpret the parameters of the multivariate normal in terms of the marginal means, variances and (bivariate) covariances, e.g.: $(\mathbf{\Sigma})_{1,2} = \mathrm{Cov}(Y_1, Y_2)$.

Multivariate normal distribution

The matrix $\mathbf{\Sigma}$ is often parameterized as:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_p\rho_{1p} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ \sigma_1\sigma_p\rho_{1p} & \cdots & \cdots & \sigma_p^2 \end{pmatrix}$$

where:

$$\begin{aligned} \sigma_j^2 &= \mathrm{Var}(Y_{ij}) \\ \rho_{j_1,j_2} &= \frac{\mathrm{Cov}(Y_{ij_1}, Y_{ij_2})}{\sqrt{\mathrm{Var}(Y_{ij_1})}\sqrt{\mathrm{Var}(Y_{ij_2})}} \end{aligned}$$

The latter is the *correlation* between Y_{ij1} and Y_{ij2}.

Multivariate normal distribution

The parameterization in matrix form:

$$\mathbf{\Sigma} = \tilde{\mathbf{\Sigma}}_d \mathbf{R} \tilde{\mathbf{\Sigma}}_d$$

where:

$$\tilde{\mathbf{\Sigma}}_d = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_p \end{pmatrix}$$

$$\mathbf{R} = \underbrace{\begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{12} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1p} & \cdots & \cdots & 1 \end{pmatrix}}_{\text{correlation matrix}}$$

Multivariate normal distribution

From covariance to correlation matrix:

$$\mathbf{R} = \tilde{\mathbf{\Sigma}}_d^{-1} \mathbf{\Sigma} \tilde{\mathbf{\Sigma}}_d^{-1}$$

where:

$$\tilde{\mathbf{\Sigma}}_d^{-1} = \begin{pmatrix} \sigma_1^{-1} & 0 & \cdots & 0 \\ 0 & \sigma_2^{-1} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_p^{-1} \end{pmatrix}$$

Question

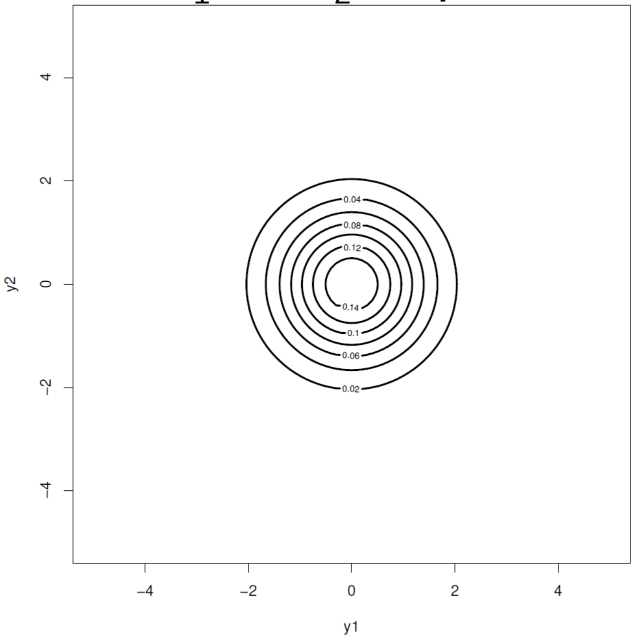
→ Verify for p=2.

→ How to go from correlation to covariance matrix?

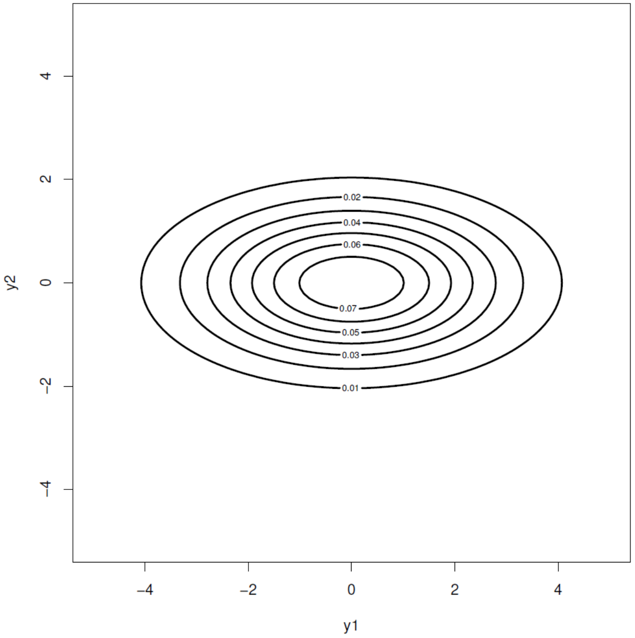
Multivariate normal distribution

Effect of σ_1, σ_2, ρ in the bivariate normal distribution.

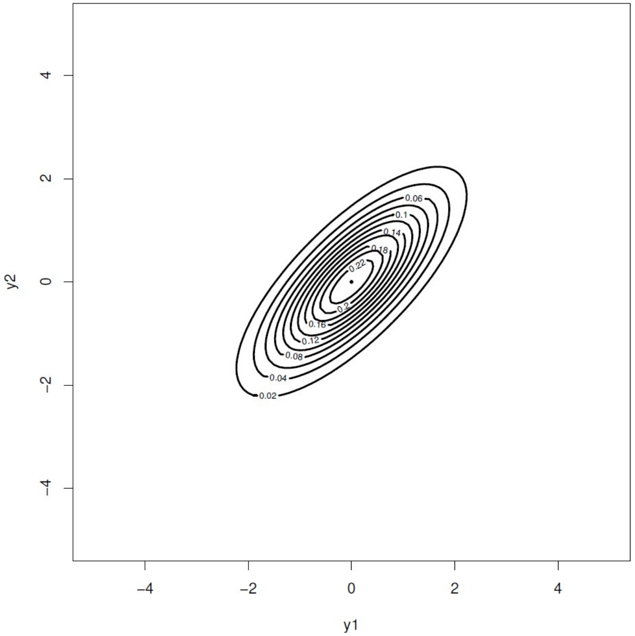
$\sigma_1=1$, $\sigma_2=1$, $\rho=0$



$\sigma_1=2$, $\sigma_2=1$, $\rho=0$



$\sigma_1=1$, $\sigma_2=1$, $\rho=3/4$



Multivariate normal distribution

Independence

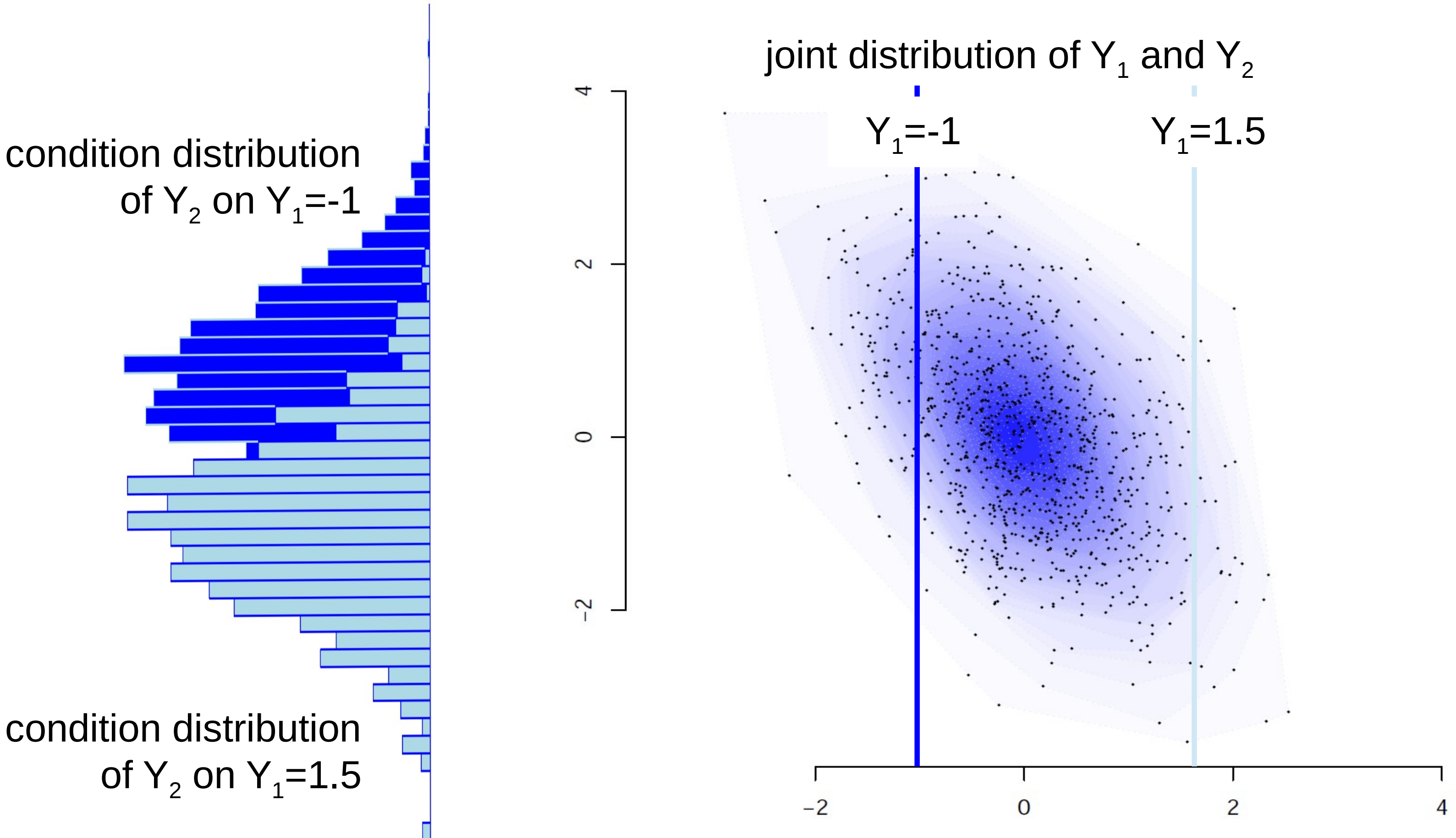
Suppose ρ = 0. Then:

$$
\begin{aligned}
& f(Y_1, Y_2) \\
= \quad & C \exp\left[-\frac{1}{2}\begin{pmatrix} Y_1-\mu_1 \\ Y_2-\mu_2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}^{-1} \begin{pmatrix} Y_1-\mu_1 \\ Y_2-\mu_2 \end{pmatrix}\right] \\
= \quad & C \exp\left[-\frac{1}{2\sigma_1^2}(Y_1-\mu_1)^2 - \frac{1}{2\sigma_2^2}(Y_2-\mu_2)^2\right] \\
= \quad & C \exp\left[-\frac{1}{2\sigma_1^2}(Y_1-\mu_1)^2\right] \exp\left[-\frac{1}{2\sigma_2^2}(Y_2-\mu_2)^2\right] \\
= \quad & g(Y_1)\, h(Y_2)
\end{aligned}
$$

Hence, the genes in a two-gene pathway are independent if ρ = 0.

Multivariate normal distribution

Partition a *p*-variate normal random variable into two exhaustive, exclusive subsets. The *conditional distribution* of a subset of variates conditioned on the other is then normally distributed.



condition distribution of Y_2 on Y_1=-1

condition distribution of Y_2 on Y_1=1.5

joint distribution of Y_1 and Y_2

Y_1=-1

Y_1=1.5

Multivariate normal distribution

Formally, assume the partitioned random variable:

$$\begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \right)$$

Theorem 6.5 of Bickel & Doksum (2001) then states:

$$\begin{aligned} \mathbf{Y}_a | \mathbf{Y}_b \quad \sim \quad & \mathcal{N}[\boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{Y}_b - \boldsymbol{\mu}_b), \\ & \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}] \end{aligned}$$

Note:

→ The theorem saves nasty integrals.

→ Joint, marginal and conditional distributions are normal.

→ The condition variance does not depend on $\mathbf{Y}_b$.

Multivariate normal distribution

Example

Consider the trivariate normal distribution:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -1 & -1 \\ -1 & 3/2 & 1/2 \\ -1 & 1/2 & 3/2 \end{pmatrix} \right)$$

Calculate the distribution of (Y_2, Y_3) conditional on Y_1.

Set $A = \{2, 3\}$ and $B = \{1\}$ and apply the Theorem from the previous slide. For the conditional mean, we obtain:

$$\boldsymbol{\mu}_{a|b} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \end{pmatrix} \cdot \frac{1}{2} \cdot (Y_1 - 0)$$

Multivariate normal distribution

Example (continued)

The conditional variance is:

$$\mathbf{\Sigma}_{a|b} = \begin{pmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{pmatrix} - \begin{pmatrix} -1 \\ -1 \end{pmatrix} \cdot \frac{1}{2} \cdot \begin{pmatrix} -1 & -1 \end{pmatrix}$$

The distribution of (Y_2, Y_3) conditional on Y_1 is thus:

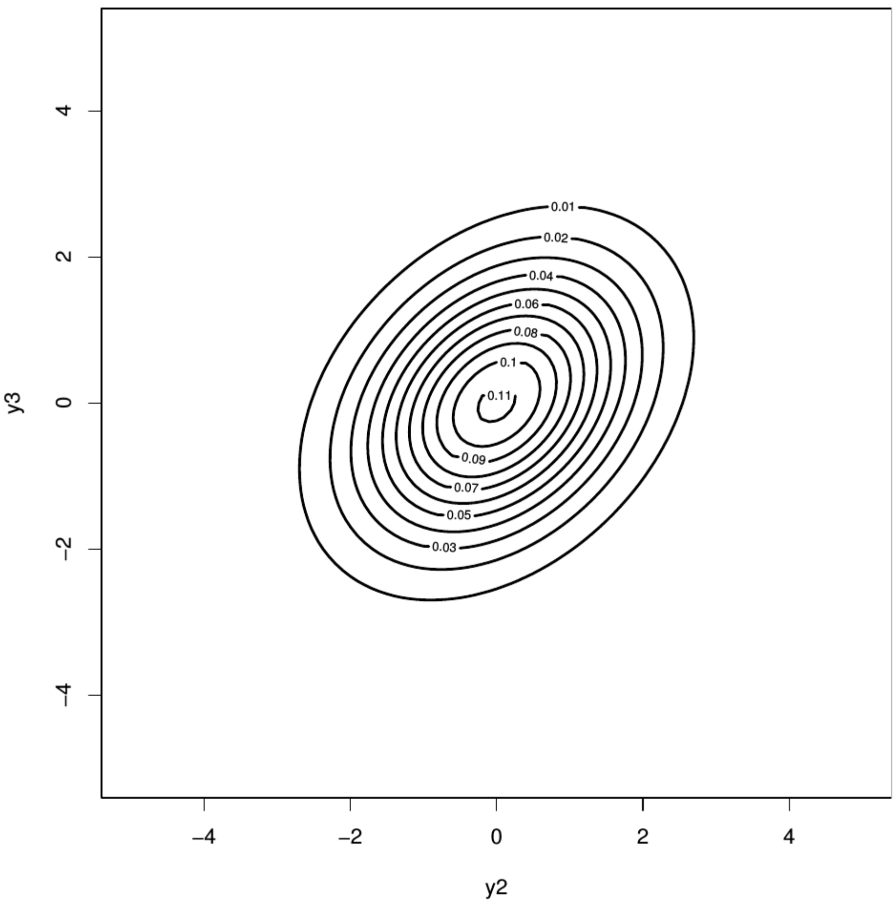
$$\begin{pmatrix} Y_2 \\ Y_3 \end{pmatrix} \Bigg| Y_1 \sim \mathcal{N}\left(\begin{pmatrix} -Y_1/2 \\ -Y_1/2 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

Hence, conditional on Y_1, variables Y_2 and Y_3 are uncorrelated.

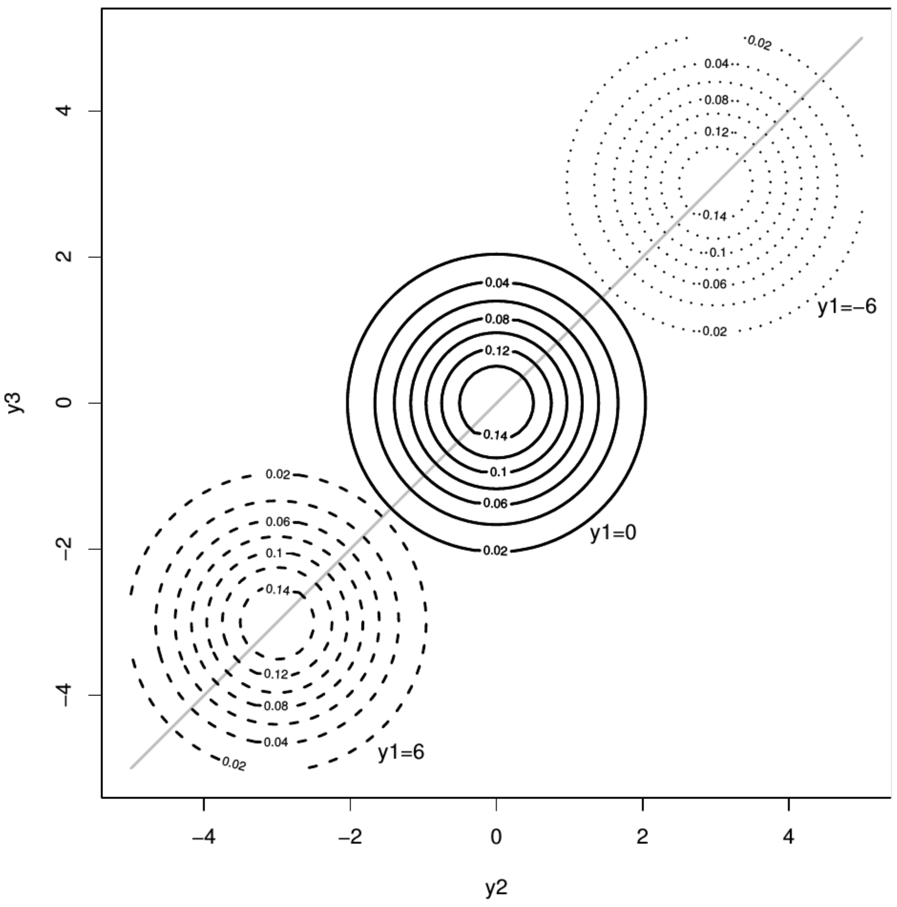
Compare this to the marginal distribution: $Y_2 \sim \mathcal{N}(0, 3/2)$

Multivariate normal distribution

Example (continued)



Marginal distribution



Conditional distribution

Multivariate normal distribution

Example

Suppose the expression levels of gene B are determined by that of gene A and some noise. That is, Y_b is the sum of two random variables:

$$Y_b \quad = \quad Y_a + \varepsilon$$

Furthermore, Y_a and ε are independent and both normally distributed with mean zero and unit variance:

$$\begin{pmatrix} Y_a \\ \varepsilon \end{pmatrix} = \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

Question

What are the mean and variance of Y_b?

And, the mean and variance of Y_b conditional on Y_a?

Multivariate normal distribution

Example

With respect to the mean:

$$\begin{aligned} E(Y_b) &= E(Y_a + \varepsilon) \\ &= E(Y_a) + E(\varepsilon) \;=\; 0 + 0 \;=\; 0 \end{aligned}$$

where the independence between Y_a and ε has been used.

Alternatively (using the zero mean of Y_a and ε):

$$\begin{aligned} E(Y_b) &= E(Y_a + \varepsilon) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_a + \varepsilon)\, \underbrace{f(y_a, \varepsilon)}_{\text{joint density}}\, \mathrm{d}y_a\, \mathrm{d}\varepsilon \end{aligned}$$

Multivariate normal distribution

Example

With respect to the variance:

$$\begin{aligned}\mathrm{Var}(Y_b) &= \mathrm{Var}(Y_a + \varepsilon) \\ &= \mathrm{Var}(Y_a) + \mathrm{Var}(\varepsilon) + 2\mathrm{Cov}(Y_a, \varepsilon) \\ &= 1 + 1 + 0 \;=\; 2\end{aligned}$$

again using the independence between Y_a and ε.

Alternatively (using the zero mean of Y_a and ε):

$$\begin{aligned}\mathrm{Var}(Y_b) &= \mathrm{Var}(Y_a + \varepsilon) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (y_a + \varepsilon)^2 \, f(y_a, \varepsilon) \, \mathrm{d}y_a \, \mathrm{d}\varepsilon\end{aligned}$$

Multivariate normal distribution

Example

With respect to the conditional mean, conditioning on Y_a means that Y_a is no longer random but fixed at some value y_a. This propagates through the calculation:

$$
\begin{aligned}
&E(Y_b \,|\, Y_a = y_a) \\
&\quad = \quad E(Y_a + \varepsilon \,|\, Y_a = y_a) \\
&\quad = \quad E(Y_a \,|\, Y_a = y_a) + E(\varepsilon \,|\, Y_a = y_a) \\
&\quad = \quad E(y_a) + E(\varepsilon) \;\; = \;\; y_a + 0 \;\; = \;\; y_a
\end{aligned}
$$

Multivariate normal distribution

Example

With respect to the conditional variance:

$$
\begin{aligned}
&\mathrm{Var}(Y_b \,|\, Y_a = y_a) \\
&\quad = \quad \mathrm{Var}(Y_a + \varepsilon \,|\, Y_a = y_a) \\
&\quad = \quad \mathrm{Var}(Y_a \,|\, Y_a = y_a) + \mathrm{Var}(\varepsilon \,|\, Y_a = y_a) \\
&\qquad\qquad +2\,\mathrm{Cov}(Y_a, \varepsilon \,|\, Y_a = y_a) \\
&\quad = \quad \mathrm{Var}(y_a) + \mathrm{Var}(\varepsilon) + 2\,\mathrm{Cov}(y_a, \varepsilon) \\
&\quad = \quad 0 + 1 + 0 \quad = \quad 1
\end{aligned}
$$

Estimation

Parameter estimation

Let **Y**$_1$, …, **Y**$_n$ be *p*-dimensional, normally distributed, random variables with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\mathbf{Y}_i \quad \sim \quad \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

The ML parameter estimates are then:

$$\hat{\boldsymbol{\mu}} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_i \quad = \quad \frac{1}{n}\left(\sum_{i=1}^{n}Y_{i1}, \ldots, \sum_{i=1}^{n}Y_{ip}\right)$$

and

$$\hat{\boldsymbol{\Sigma}} \quad = \quad \mathbf{S}$$

where

$$\mathbf{S} \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\mu})\,(\mathbf{Y}_i - \boldsymbol{\mu})^T$$

Estimation

Parameter estimation

These estimates are the standard univariate estimators aggregated into vector and matrix:

$$(\hat{\boldsymbol{\mu}})_j = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i \right)_j = \frac{1}{n} \sum_{i=1}^{n} Y_{i,j} = \hat{\mu}_j$$

$$(\hat{\boldsymbol{\Sigma}})_{j,j} = (\mathbf{S})_{j,j} = \left(\frac{1}{n} \sum_{i=1}^{n} (\mathbf{Y}_i - \boldsymbol{\mu})(\mathbf{Y}_i - \boldsymbol{\mu})^\top \right)_{j,j}$$

$$= \frac{1}{n} \sum_{i=1}^{n} (Y_{i,j} - \mu_j)^2 = \hat{\sigma}_j^2$$

and similarly for the off-diagonal elements of the covariance matrix.

Why the multivariate normal distribution?

(supplementary material)

Why multivariate normal?

Motivation from rate equations

The transcriptional process is often modeled by rate equations, a system of ordinary differential equations.

The rate equations model the regulatory process by linking a change (over time) in one gene’s transcripts to the mRNA concentrations of the other genes in the pathway:

$$\underbrace{\frac{d\,Y_1}{d\,t}}_{\text{change over time}} = \underbrace{f(Y_1, \ldots, Y_p)}_{\text{transcription}} - \underbrace{\gamma\,Y_1}_{\text{degradation}}$$

Why multivariate normal?

Motivation from rate equations

$$
\begin{array}{ccccc}
dY_{i1}/dt & = & f_1(Y_{i1},\ldots,Y_{ip}) & - & \gamma_1\,Y_{i1} \\
\vdots & & \vdots & & \vdots \\
\underbrace{dY_{ip}/dt}_{\substack{\text{production}\\ \text{rate}}} & = & \underbrace{f_p(Y_{i1},\ldots,Y_{ip})}_{\text{transcription}} & - & \underbrace{\gamma_p\,Y_{ip}}_{\substack{\text{degra-}\\ \text{dation}}}
\end{array}
$$

Why multivariate normal?

Motivation from rate equations

$$\begin{array}{ccccc} 0 & = & f_1(Y_{i1},\ldots,Y_{ip}) & - & \gamma_1 \, Y_{i1} \\ \vdots & & \vdots & & \vdots \\ \underbrace{0}_{\substack{\text{production}\\ \text{rate}}} & = & \underbrace{f_p(Y_{i1},\ldots,Y_{ip})}_{\text{transcription}} & - & \underbrace{\gamma_p \, Y_{ip}}_{\substack{\text{degra-}\\ \text{dation}}} \end{array}$$

Assumption 1: steady state

Why multivariate normal?

Motivation from rate equations

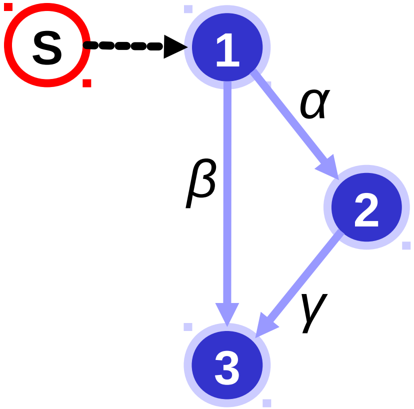
$$\begin{array}{ccccc} 0 & = & \sum_{j=1}^{p} \theta_{1j} \, Y_{ij} & - & \gamma_1 \, Y_{i1} \\ \vdots & & \vdots & & \vdots \\ \underbrace{0}_{\substack{\text{production}\\ \text{rate}}} & = & \underbrace{\sum_{j=1}^{p} \theta_{pj} \, Y_{ij}}_{\text{transcription}} & - & \underbrace{\gamma_p \, Y_{ip}}_{\substack{\text{degra-}\\ \text{dation}}} \end{array}$$

Assumption 1: steady state

Assumption 2: linearize

Finally, assume the *Y*’s random and add error term.

Why multivariate normal?



Motivation from rate equations

Original

$$
\begin{aligned}
0 &= && - Y_1 \\
0 &= \alpha Y_1 && - Y_2 \\
0 &= \beta Y_1 + \gamma Y_2 && - Y_3
\end{aligned}
$$

New

$$
\begin{aligned}
Y_1 &= && S + e_3 \\
Y_2 &= \alpha Y_1 && + e_3 \\
Y_3 &= \beta Y_1 + \gamma Y_2 && + e_3
\end{aligned}
$$

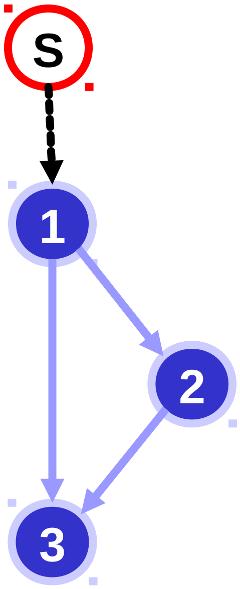
Why multivariate normal?

Why the multivariate normal?

Consider a pathway of 3 genes.

Assuming the expression of the genes in the pathway follows a linear system:

$$\begin{aligned} Y_1 &= S + \varepsilon_1 \\ Y_2 &= \alpha Y_1 + \varepsilon_2 \\ Y_3 &= \beta Y_1 + \gamma Y_2 + \varepsilon_3 \end{aligned}$$



with the signal and errors independent and normal.

As the sum of normally distributed variables is also normally distributed, all genes are normally distributed!

Why multivariate normal?

Calculate mean and variance of resulting trivariate normal distribution by means of expectation and variance rules. E.g.:

$$\begin{aligned}
E(Y_1) &= E(S) + E(\varepsilon_1) \\
\mathrm{Var}(Y_1) &= \mathrm{Var}(S) + \mathrm{Var}(\varepsilon_1) + 2\,\mathrm{Cov}(S, \varepsilon_1)
\end{aligned}$$

$$\begin{aligned}
E(Y_2) &= \alpha\, E(Y_1) + E(\varepsilon_2) \\
\mathrm{Var}(Y_2) &= \alpha^2\, \mathrm{Var}(Y_1) + \mathrm{Var}(\varepsilon_2) + 2\,\mathrm{Cov}(Y_1, \varepsilon_2) \\
\mathrm{Cov}(Y_1, Y_2) &= \mathrm{Cov}(Y_1, \alpha\, Y_1 + \varepsilon_2) \\
&= \alpha\, \mathrm{Cov}(Y_1,\ Y_1) + \mathrm{Cov}(Y_1, \varepsilon_2)
\end{aligned}$$

This is generalized in the next theorem.

Why multivariate normal?

Theorem (Koller, Friedman, 2009)

Suppose $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$, $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, and define:

$$Y = \beta_0 + \boldsymbol{\beta}^T \mathbf{X} + \varepsilon$$

Then, $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ with:

$$\begin{aligned} \mu_Y &= \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\mu}_X \\ \sigma_Y^2 &= \sigma^2 + \boldsymbol{\beta}^T \boldsymbol{\Sigma}_X \boldsymbol{\beta} \end{aligned}$$

and

$$\mathrm{Cov}(X_{j_1}, Y) = \sum_{j_2=1}^{p} \beta_{j_2} (\boldsymbol{\Sigma}_X)_{j_1, j_2}$$

Why multivariate normal?

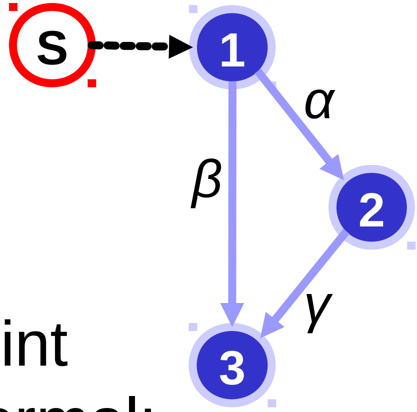


Illustration of theorem

The last theorem enables the calculation of the joint distribution of Y_1, Y_2, and Y_3. It is a multivariate normal:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ * \\ * \end{pmatrix}, \begin{pmatrix} \sigma^2_{Y_1} & * & * \\ * & * & * \\ * & * & * \end{pmatrix} \right)$$

The theorem tells us how to fill the gaps:

$$\begin{aligned} \mu_{Y_2} &= \beta_0 + \alpha \mu_{Y_1} \\ \sigma^2_{Y_2} &= \sigma^2_{\varepsilon_2} + \alpha^2 \sigma^2_{Y_1} \\ \mathrm{Cov}(Y_1, Y_2) &= \alpha \sigma^2_{Y_1} \end{aligned}$$

Why multivariate normal?

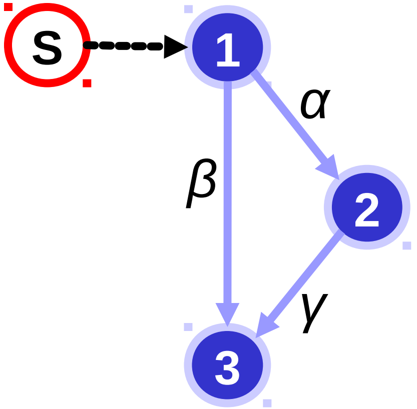


Illustration of theorem

So far, we thus have:

$$\mathcal{N}\left(\left(\begin{array}{c} 0 \\ 0 \\ * \end{array}\right), \left(\begin{array}{ccc} \sigma_{Y_1}^2 & \alpha\sigma_{Y_1}^2 & * \\ \alpha\sigma_{Y_1}^2 & \sigma_{\varepsilon_2}^2 + \alpha^2\sigma_{Y_1}^2 & * \\ * & * & * \end{array}\right)\right)$$

The rest goes in a similar fashion, e.g.:

$$\sigma_{Y_3}^2 \quad = \quad \sigma_{\varepsilon_3}^2 + \left(\begin{array}{c} \beta \\ \gamma \end{array}\right)^T \left(\begin{array}{cc} \sigma_{Y_1,Y_1} & \sigma_{Y_1,Y_2} \\ \sigma_{Y_2,Y_1} & \sigma_{Y_2,Y_2} \end{array}\right) \left(\begin{array}{c} \beta \\ \gamma \end{array}\right)$$

Why multivariate normal?

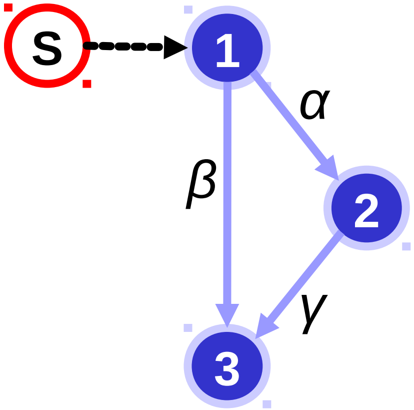


Illustration of theorem

Finally, this gives:

$$\boldsymbol{\mu} \quad = \quad (\mu_{Y_1}, \mu_{Y_2}, \mu_{Y_3})^T \quad = \quad (0,0,0)^T$$

and $\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}[(Y_1, Y_2, Y_3)^T]$

$$\begin{pmatrix} \sigma_{Y_1}^2 & \alpha\sigma_{Y_1}^2 & \delta\sigma_{Y_1}^2 \\ \alpha\sigma_{Y_1}^2 & \alpha^2\sigma_{Y_1}^2 + \sigma_{\varepsilon_2}^2 & \alpha\delta\sigma_{Y_1}^2 + \gamma\sigma_{\varepsilon_2}^2 \\ \delta\sigma_{Y_1}^2 & \alpha\delta\sigma_{Y_1}^2 + \gamma\sigma_{\varepsilon_2}^2 & \delta^2\sigma_{Y_1}^2 + \gamma^2\sigma_{\varepsilon_2}^2 + \sigma_{\varepsilon_3}^2 \end{pmatrix}$$

where $\delta = \beta + \alpha\gamma$

cc creative commons

BY: =