Learning about probabilistic inference and forecasting playing with multivariate normal distributions (with examples in R)

G. D'Agostini
Università "La Sapienza" and INFN, Roma, Italia
(giulio.dagostini@roma1.infn.it, http://www.roma1.infn.it/~dagos)

March 4, 2014

Abstract

The properties of the normal distribution under linear transformation, as well the easy way to compute the covariance matrix of marginals and conditionals, offer a unique opportunity to get an insight about several aspects of uncertainties in measurements. The way to build the initial covariance matrix in a few, but concettually relevant cases is illustrated: several observations made with (possibly) different instruments measuring the same quantitiy; effect of systematics (although limited to *offset*, in order to stick to linear models) on the determination of the 'true value' as well in the prediction of future observations; correlations which arise when different quantities are measured with the same instrument affected by an offset uncertainty. Many numerical examples are provided, exploiting the ability of the R language to handle large matrices and to produce high quality plots. Some of the results will be framed in the general problem of 'propagation of evidence', crucial in analysing graphical models of knowledge.

"So far as the theories of mathematics are about reality, they are not certain; so far as they are certain, they are not about reality.
(A. Einstein)

"If we were not ignorant there would be no probability, there could only be certainty. But our ignorance cannot be absolute, for then there would be no longer any probability at all."
(H. Poincaré)

"Probability is good sense reduced to a calculus"
(S. Laplace)

"All models are wrong but some are useful"
(G. Box)

1 Introduction

The opening quotes set up the frame in which this paper has been written: in the sciences we always deal with uncertainties; being in condition on uncertainty we can only state 'somehow' how much we believe something; in order to do that we need to build up probabilistic models based on good sense. For example, if we are uncertain about the value we are going to *read on* an instrument, we can make probabilistic assessments about it. But in general our interest is the *numerical value of a physics quantity*. We are usually in great condition of uncertainty before the measurement, but we still remain with some degree of uncertainty after the measurement has been performed. Models enter in the construction of the the causal network which connects physics quantities to what we can observe on the instruments. They are also important because it is convenient to use, whenever it is possible, probability distributions, instead than to assign individual probabilities to each individual 'value' (after suitable discretization) that a physics quantity might assume.

As we know, there are good reasons why in many cases the Gaussian distribution (or *normal* distribution) offers a *reasonable* and *convenient* description of the probability that the quantity of interest lies within some bounds. But it is important to remember that, as it was clear to Gauss [2] when he derived the famous distribution for the measurement errors, one should not take literally the fact that the variable appearing in the formula can range from minus infinite to plus infinite: an apple cannot have infinite mass, or a negative one!

Sticking hereafter to Gaussian distributions, it is clear that if we are only interested to the probability density function (pdf) of a variable at the time, we can only describe our uncertainty about that quantity, and nothing more. The interesting thing is when we study the joint distribution of several variables, because this is the way we can learn about some of them assuming the values of the others. For example, if we assume the joint pdf $f(x_1, x_2 \,|\, I)$ of variables X_1 and X_2 under the state of information I (on which we ground our assumptions), we can evaluate $f(x_1 \,|\, x_2\,, I)$, that is the pdf adding the extra condition $X_2 = x_2$, which is usually not the same as $f(x_1 \,|\, I)$, the pdf of X_1 for any value X_2 might assume.[1]

Let us take for example the three diagrams of Fig. 1 to which we give a physical interpretation:

1. In the diagram on the left the variable X_1 might represent the numerical value of a physics quantity, on which we are in condition on uncertainty, modelled by

$$X_1 \sim \mathcal{N}(X_0, \sigma_1)\,, \tag{1}$$

[1]The pdf $f(x_1 \,|\, I)$ is called *marginal*, although there is never special about this name, since all distributions of a single variable can be thought as being 'marginal' to all other possible quantities which we are not interested about. $f(x_1 \,|\, x_2\,, I)$ is instead 'called' *conditional*, although it is a matter of fact that all distributions are conditional to a given state of information, here indicated by I. Note that throughout this paper will shall use the same symbol $f()$ for all pdf's, as it is customary among physicists – I have met mathematics oriented guys getting mad by the equation $f(x, y) = f(x|y) \cdot f(y)$ because, they say, "the three functions cannot be the same"...

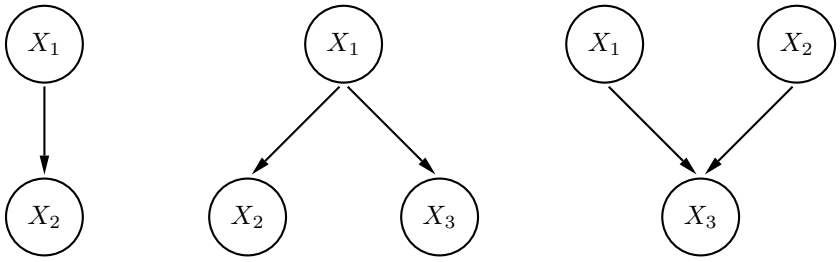


Figure 1: Basic models of joint probabilities

where X_0 and σ_1 are suitable parameters to state our 'ignorance' about X_1 ('complete ignorance', if it does ever exist, is recovered in the limit $\sigma_1 \to \infty$). Instead, X_2 is then what we read on an instrument when we apply to X_1. That is, even if we knew X_1, we are still uncertain about what we can read on the instrument, as well understood. Modelling this uncertainty by a normal distribution we have, for any value of X_1

$$X_2|_{X_1} \sim \mathcal{N}(X_1, \sigma_{2|1})\,, \tag{2}$$

where $\sigma_{2|1}$ is a compact symbol for $\sigma(X_2|_{X_1})$ and which is in general different from $\sigma_2 \equiv \sigma(X_2)$. In fact our uncertainty about X_2 (for any possible value of X_1) must be larger than that about X_1 itself, for obvious reasons – which shall see later the details.

2. In the diagram on the center X_3 might represent a second observation done *independently* applying in general a second instrument to the identical value X_1. This means that $X_2|_{X_1}$ and $X_3|_{X_1}$ are independent, although X_2 and X_3 are <u>not</u>, as we shall see.

3. In the diagram on the right X_3 is the observation read on the instrument applies to X_1, but possibly influenced by X_2, that might then represent a kind of *systematics*.

Note, how it has been precisely stated, that X_2 of the first and of the second diagrams, as well as X_3 of the other two, are the *readings* on the instruments and <u>not</u> the result of the measurement! This is because by "result of the measurement" we mean statements about the quantity of interest and not about the quantities read on the instruments (think for example at the an experiment measuring the Higgs boson mass, making use of the information recorded by the detector!). In this case the "result of the measurement" would be $f(x_1 \,|\, \mathbf{data}\,, I)$ where **data** stands for the set of observed variables.

The diagrams of the figure can be complicated, using sets of data, with systematics effect common to observations in each subset. The aim of this paper is to help in developing some intuition of what is going on in problems of this kind, with the only simplification that all pdf's of interest will be normal.

2 Technical premises (with some exercises)

We assume that the reader is familiar with some basic concepts related to *uncertain numbers* and *uncertain vectors*, usually met under the name of "random variables".

2.1 Normal (Gaussian) distribution

$X \sim \mathcal{N}(\mu, \sigma)$:

$$f(x\,|\,\mathcal{N}(\mu,\sigma)) = \frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp\left[-\frac{(x-\mu)^2}{2\,\sigma^2}\right] \tag{3}$$

with

$$\mathrm{E}[X] = \mu \tag{4}$$

$$\mathrm{Var}[X] = \sigma^2 \tag{5}$$

$$\sigma[X] = \sqrt{\mathrm{Var}[X]} = \sigma\,. \tag{6}$$

(We remind that in most physics applications $x \to \pm\infty$ simply means $|x-\mu|/\sigma \gg 1$.)

In R [1] there are functions to calculate the pdf ('`dnorm()`'), the cumulative function, usually indicated with "$F(x)$" ('`pnorm()`') as well as its inverse ('`qnorm()`'), as easily shown in the following examples[2] ('>' is the R console prompt):

```
> dnorm(0, 0, 1)
[1] 0.3989423
> 1/sqrt(2*pi)   # (just a check)
[1] 0.3989423
> pnorm(0, 0, 1)
[1] 0.5
> pnorm(7, 5, 2) - pnorm(3, 5, 2)
[1] 0.6826895
> qnorm(0.5, 5, 2)
[1] 5
> qnorm(1, 5, 2)
[1] Inf
> qnorm(0, 5, 2)
[1] -Inf
```

Note the capability of the language to handle infinities, as it can be cross checked by

```
> pnorm(Inf, 5, 2)
[1] 1
```

And here are the instructions to produce the plots of figure 2.

```
mu <- 5; sigma <- 2; x <- seq(mu-5*sigma, mu+5*sigma, len=101)
plot(x, dnorm(x, mu, sigma), ty='l', ylab='f(x)', col='blue')
points(x, dnorm(x, mu, sigma*1.5), ty='l', lty=2, col='blue')
points(x, dnorm(x, mu, sigma*2), ty='l', lty=3, col='blue')
```

[2] For information about the language se one of the many tutorial available on the web. Most functions we shall use here have self explaining names. For an help, for example about `dnorm()`, just enter

```
> ?dnorm
```

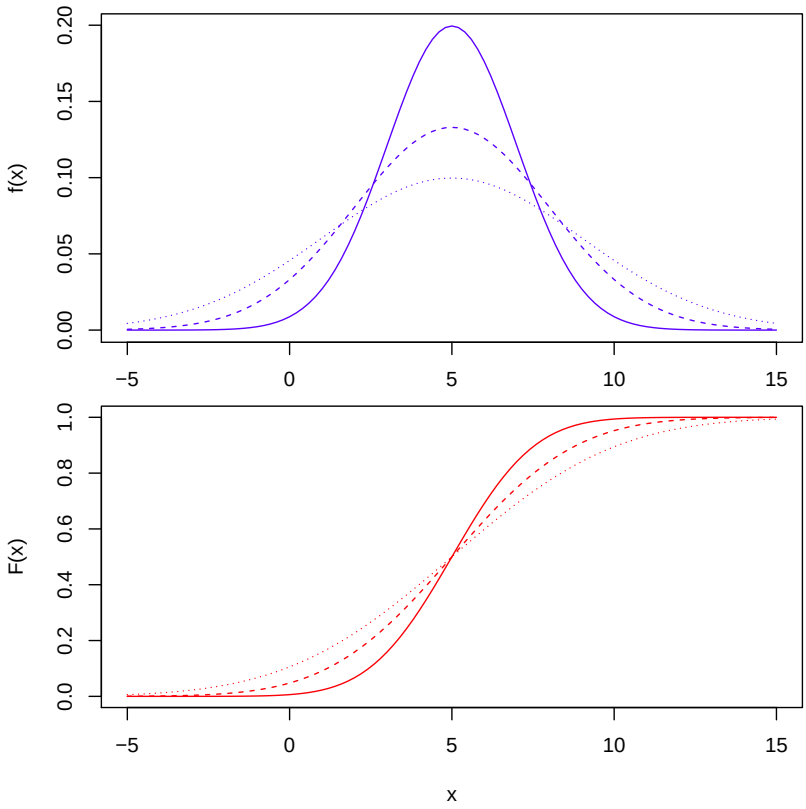


Figure 2: Gaussian probability density function (above) and cumulative function (below) for $\mu = 5$ and $\sigma = 2$, 3 and 4 (solid, dashed and pointed).

```
plot(x, pnorm(x, mu, sigma), ty='l', ylab='F(x)', col='red')
points(x, pnorm(x, mu, sigma*1.5), ty='l', lty=2, col='red')
points(x, pnorm(x, mu, sigma*2), ty='l', lty=3, col='red')
```

2.2 Bivariate and multivariate normal distribution

A bivariate normal distribution is defined by

$$f(\boldsymbol{x}\,|\,\mathcal{N}(\boldsymbol{\mu},\boldsymbol{V})) = \frac{1}{2\,\pi\,\sigma_1\,\sigma_2\,\sqrt{1-\rho_{12}^2}}\,\exp\left\{-\frac{1}{2\,(1-\rho_{12}^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\,\rho_{12}\,\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\,\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\}, \tag{7}$$

where

$$\begin{aligned} \boldsymbol{x} &= (x_1, x_2) && (8)\\ \boldsymbol{\mu} &= (\mu_1, \mu_2) && (9)\\ \mathrm{E}[X_i] &= \mu_i && (10)\\ \mathrm{Var}[X_i] &= \sigma_i^2 && (11)\\ \sigma[X_i] \equiv \sqrt{\mathrm{Var}[X_i]} &= \sigma_i && (12)\\ \rho_{12} &= \frac{\mathrm{Cov}[X_1, X_2]}{\sigma_1\,\sigma_2}\,, && (13) \end{aligned}$$

with variances and covariances forming the *covariance matrix*

$$\boldsymbol{V} = \begin{pmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] \\ \mathrm{Cov}[X_1, X_2] & \mathrm{Var}[X_2] \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho_{12}\,\sigma_1\,\sigma_2 \\ \rho_{12}\,\sigma_1\,\sigma_2 & \sigma_2^2 \end{pmatrix} \qquad (14)$$

The bivariate pdf (7) can be rewritten in a compact form as

$$f(\boldsymbol{x}\,|\,\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{V})) = (2\pi)^{-n/2} |\boldsymbol{V}|^{-1/2} \exp\left[-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu})^T\,\boldsymbol{V}^{-1}\,(\boldsymbol{x}-\boldsymbol{\mu})\right], \qquad (15)$$

expression valid for any number of variables. ($|\boldsymbol{V}|$ stands for $\det(\boldsymbol{V})$.)

2.2.1 Multivariate normals in R

Functions to calculate multivariate normal pdf's, as well as cumulative functions and random generators are provided in R via the package `mnormt`[3] that needs to be installed[4] and then loaded by the command

```
> library(mnormt)
```

Then we have to define the values of the parameters and built up the vector of the central values and the covariance matrix. Here is an example:

```
> m1=0.4; m2=2; s1=1; s2=0.5; rho=0.6
> mu <- c(m1, m2)
> ( V <- rbind( c( s1^2, rho*s1*s2), c(rho*s1*s2, s2^2) ) )
     [,1] [,2]
[1,] 1.0 0.30
[2,] 0.3 0.25
```

Then we can evaluate the joint pdf in a point (x_1, x_2), e.g.

```
> dmnorm(c(0.5, 1.5), mu, V)
```

[3] `http://cran.r-project.org/web/packages/mnormt/`

[4] For all technical details about R (open source and multi-platform!) see the R web site [1]. For example, the command to install `mnormt` is

```
> install.packages("mnormt")
```

```
[1] 0.1645734
```
Or we can evaluate $P(X_1 \leq 0.5 \;\&\; X_2 \leq 1.5)$, or $P(X_1 \leq \mu_1 \;\&\; X_2 \leq \mu_2)$, respectively, with
```
> pmnorm(c(0.5, 1.5), mu, V)
[1] 0.140636
```
and
```
> pmnorm(mu, mu, V)
[1] 0.3524164
```

2.3 Graphical representation of normal bivariates

If we like to visualize the joint distribution we need a 3D graphical package, for example `rgl`[5] or `plot3D`.[6] We need to evaluate the joint pdf on a grid of values 'x' and 'y' and provide them to the function `persp3d()`. Here are the instructions that use the `rgl` package:
```
> library(rgl)
> fun <- function(x1,x2) dmnorm(cbind(x1, x2), mu, V)
> x1 <- seq(m1-3*s1, m1+3*s1, len=51)
> x2 <- seq(m2-3*s2, m2+3*s2, len=51)
> f <- outer(x1, x2, fun)
> persp3d(x1, x2, f, col='cyan', xlab="x1", ylab="x2", zlab="f(x1,y2)")
```
After the plot is shown in the graphics window, the window can be enlarged and plot rotated at wish. Figure 3 shows in the uppre two plots two views of the same distribution.

Here also the instructions to use `plot3D()`:
```
> library(plot3D)
> M <- mesh(x1, x2)
> surf3D(M$x, M$y, f, bty='b2', phi = 30, theta = -20,
+ xlab='x1', ylab='x2', zlab='f(x1,x2)')
```
The result is shown in the lower plot of Fig. 3.

Another convenient and often representation of normal bivariates is to draw iso-pdf contours, i.e. lines in correspondence of the points in the plane (x_1, x_2) such as $f(x_1, x_2 \,|\, I) =$ const. This requires that the *quadratic form* at the exponent (that is what is written in general as $(\boldsymbol{x}-\boldsymbol{\mu})^T \, \boldsymbol{V}^{-1} \, (\boldsymbol{x}-\boldsymbol{\mu})$) has a fixed value. In the two dimensional case we recognize the expression of an ellipse. We have in R the convenient package `ellipse`[7] to evaluate the points of such an ellipse, given the vector of expected matrix, the covariance matrix and the probability that a point falls inside it. Here is the script that applies it to the same bivariate normal of Fig. 3, producing the contour plots of Fig. 4

```
plot( ellipse(V, centre=mu, level=0.9973), ty='l', lty=2, col='red',
     asp=1, xlab=expression(x[1]), ylab=expression(x[2]) )
points( ellipse(V, centre=mu, level=0.99), ty='l', col='blue')
points( ellipse(V, centre=mu, level=0.954), ty='l', lty=2, col='red')
```

[5] `https://r-forge.r-project.org/projects/rgl/`
[6] http://www.r-bloggers.com/3d-plots-in-r/
[7] `http://cran.r-project.org/web/packages/ellipse/`

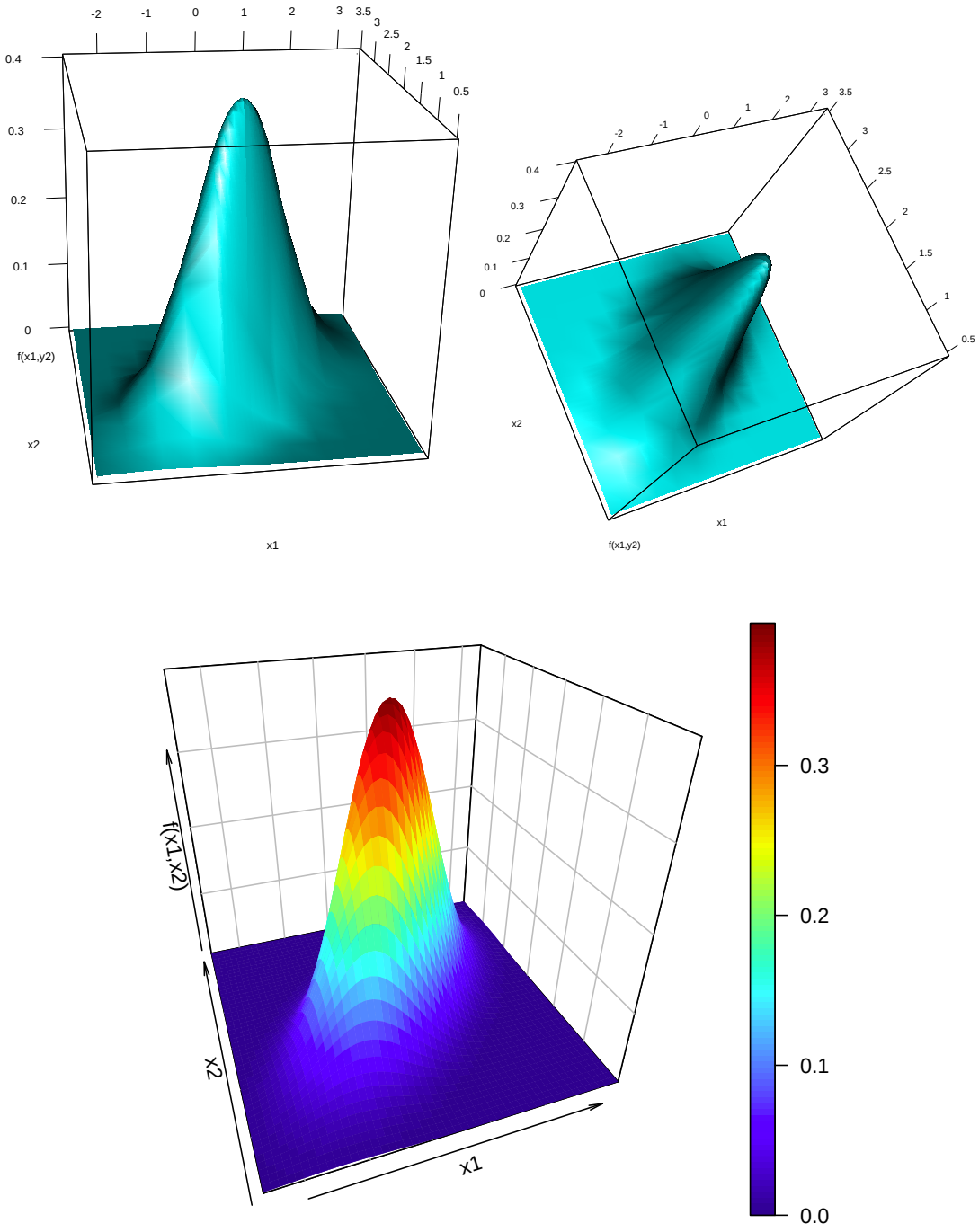


Figure 3: Three views of the bivariate normal distribution obtained with the R code provided in the text. The above two are obtained by `perp3d()` of the packege *rgl*, producing interactive 3D plots. The one below is produced by `surf3D()` of the omonymous package.

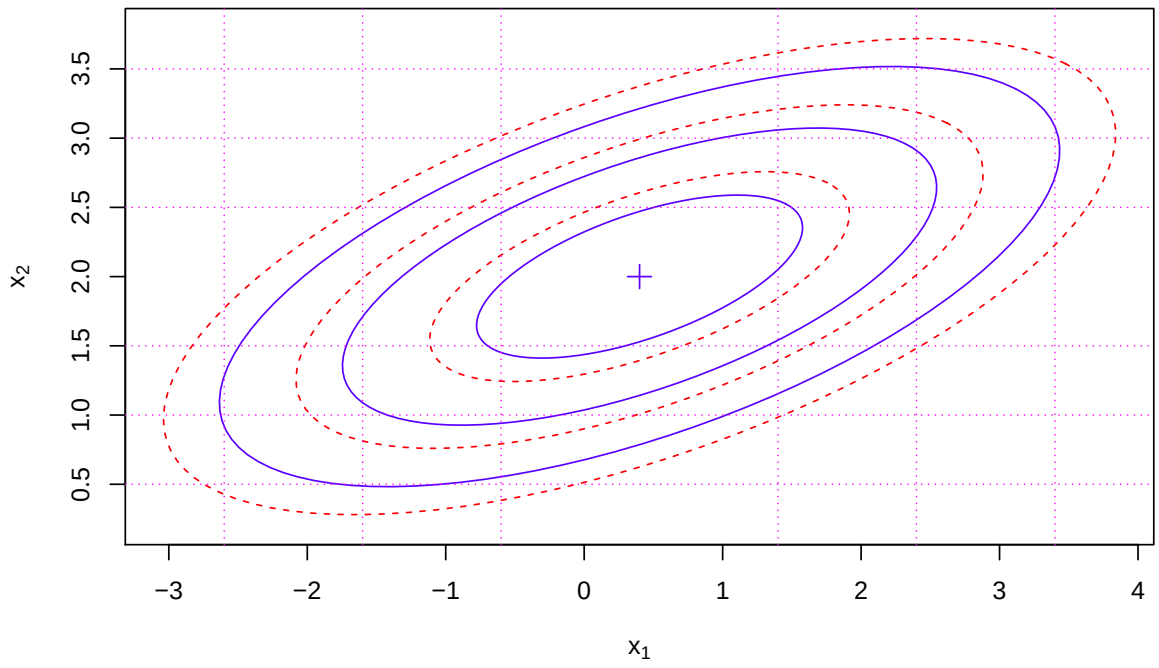


Figure 4: Contour plots of the same bivariate normal of Fig. 3. The solid lines show the ellipses inside which there is, from the smaller to the larger, 50%, 90% and 99% probability that a point x_1, x_2 could fall. The dashed ellipses define instead the 68.3%, 95.5% and 99.73% probability contours [the (in-)famous 1-σ, 2-σ and 3-σ contours, not simply related to the standard deviations of the individual variable, whose 1-σ, 2-σ and 3-σ bounds are indicated by the dotted vertical and horizontal lines].

```
points( ellipse(V, centre=mu, level=0.5), ty='l', col='blue')
points( ellipse(V, centre=mu, level=0.683), ty='l',  lty=2, col='red')
points( ellipse(V, centre=mu, level=0.90), ty='l', col='blue')
points(mu[1], mu[2], pch=3, cex=1.5, col='blue')
for(k in 1:3) {
  abline(v=mu[1]-k*sqrt(V[1,1]), lty=3, col='magenta')
  abline(v=mu[1]+k*sqrt(V[1,1]), lty=3, col='magenta')
  abline(h=mu[2]-k*sqrt(V[2,2]), lty=3, col='magenta')
  abline(h=mu[2]+k*sqrt(V[2,2]), lty=3, col='magenta')
}
```

The probability to find a point inside the ellipse contour is defined by the argument `level` (see Appendix). The ellipses drawn with solid lines define, in order of size, 50%, 90% and 99% contours. For comparison there are also the contours at 68.3%, 95.5% and 99.73%, which define the *highly confusing* 1-σ , 2-σ and 3-σ contours. Indeed, the probability that each of the variable falls in the interval of $\mathrm{E}[X_i] \pm k\,\sigma[X_i]$ has little to do with these

ellipses. And we are intersested to the probability that a point falls in a rectangles defined by $(\mathrm{E}[X_1] \pm k\,\sigma[X_1]\,\&\,\mathrm{E}[X_2] \pm k\,\sigma[X_2])$ the probability needs to be calculated making the integral of the joint distribution inside the rectangle (some of these rectangles are shown in Fig. 4 by the dotted lines, that indicate 1-σ , 2-σ and 3-σ bound in the individual variable).

Let us see how to evaluate in R the probability that a point falls in a rectangle, making use of the cumulative probability function `pmnorm()`. In fact the probability in a rectangle is related to the cumulative distribution by the following relation

$$\begin{aligned} P[\,(x_{1_m} \le X_1 \le x_{1_M})\,\&\,(x_{2_m} \le X_2 \le x_{2_M})\,] \;=\; & P[\,(X_1 \le x_{1_M})\,\&\,(X_2 \le x_{2_M})] \\ & -P[\,(X_1 \le x_{1_M})\,\&\,(X_2 \le x_{2_m})] \\ & -P[\,(X_1 \le x_{1_m})\,\&\,(X_2 \le x_{2_M})] \\ & +P[\,(X_1 \le x_{1_m})\,\&\,(X_2 \le x_{2_1})]\,, \end{aligned} \quad (16)$$

that can be implemented in an R function:

```
p.rect.norm <- function(xlim, ylim, mu, V, sigmas=FALSE, ...) {
  # the argument '...' might be useful to pass extra arguments to pmnorm

  # if sigmas is TRUE, xlim and ylim are interpreted as
  # numbers of sigmas around the mean

  if ( (length(mu) != 2) | sum( dim(V) != c(2,2) )   # some check
      | (length(xlim) != 2) | (length(ylim) != 2) ) {
    print("wrong dimensions in one of parameters")
    return(NULL)
  } else if ( sum( eigen(V)$values <= 0 ) >  0) {
    cat( sprintf("V is not positevely defined\n") )
    return(NULL)
  }

  if( sigmas) { # rectangular defined in units of individual sigma around mu
    xlim <- mu[1] + xlim * sqrt(V[1,1])
    ylim <- mu[2] + ylim * sqrt(V[2,2])
  }

  library(mnormt)
  p.rect <- pmnorm( c(xlim[2], ylim[2]), mu, V, ...) -
            pmnorm( c(xlim[2], ylim[1]), mu, V, ...) -
            pmnorm( c(xlim[1], ylim[2]), mu, V, ...) +
            pmnorm( c(xlim[1], ylim[1]), mu, V, ...)
  return(p.rect)
}
```

For example[8]

[8] For Monte Carlo oriented guys, here is how to cross check the results

```
> xy <- rmnorm(100000, mu, V)
> length( xy[,1][ xy[,1] > m1 - s1 & xy[,1] < m1+s1 & xy[,2] > m2 - s2 & xy[,2] < m2+s2 ] )
[1] 51313
```

```
> p.rect.norm(c(m1-s1, m1+s1), c(m2-s2, m2+s2), mu, V)
[1] 0.5138685
> p.rect.norm(c(-1, 1), c(-1, 1), mu, V, sigmas=TRUE)
[1] 0.5138685
```

As a cross check, let calculate the probabilities in strips of plus/minus one standard deviations around the averages:

```
> p.rect.norm(c(-1, 1), c(-10, 10), mu, V, sigmas=TRUE)
[1] 0.6826895
> p.rect.norm(c(-10, 10), c(-1, 1), mu, V, sigmas=TRUE)
[1] 0.6826895
```

2.4 Marginals (and multivariate marginals) of multivariate normals

A nice feature of the multivariate normal distribution is that if we are just interested to a subset of variables alone, neglecting which value the other ones can take ('marginalizing'), we just drop from $\boldsymbol{\mu}$ and from V the uninteresting values, or the relative rows and columns, respectively. For example, if we have – see subsection 6.1.2 –

$$\boldsymbol{\mu} = \begin{pmatrix} 1.96 \\ 0.02 \\ 1.98 \end{pmatrix} \qquad \boldsymbol{V} = \begin{pmatrix} 1.96 & -0.98 & 0.98 \\ -0.98 & 0.99 & 0.01 \\ 0.98 & 0.01 & 1.99 \end{pmatrix} \tag{17}$$

marginalizing over the second variable (i.e. being only interested in the first and the third) we obtain

$$\boldsymbol{\mu}' = \begin{pmatrix} 1.96 \\ 1.98 \end{pmatrix} \qquad \boldsymbol{V}' = \begin{pmatrix} 1.96 & 0.98 \\ 0.98 & 1.99 \end{pmatrix} \tag{18}$$

Here is a function that return espected values and variance of the multivariate 'marginal'

```
marginal.norm <- function(mu, V, x.m) {
  # x.m is vectors with logical values indicating
  # the elements on which marginalyse
  out <- NULL
  out$mu <- mu[x.m]
  v <- which(x.m)
  out$V <-  V[v, v]
  return(out)
}
```

2.5 Conditional distribution of a variable, given its bivariate distribution with another variable

A different problem is the pdf of one of variables, say X_1, for a given value of the other. This is not as straightforward as the marginal (and for this reason in this subsection we only

consider the bivariate case). Fortunately the distribution is still a Gaussian, with *shifted central value* and *squeezed width*:

$$X_1|_{x_2} \sim \mathcal{N}\left(\mu_1 + \rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(x_2-\mu_2)\,,\,\sigma_1\sqrt{1-\rho_{12}^2}\right)\,, \tag{19}$$

i.e.

$$\begin{aligned} \mathrm{E}[X_1] &= \mu_1 + \rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(x_2-\mu_2) && (20)\\ \mathrm{Var}[X_1] &= \sigma_1^2\cdot(1-\rho_{12}^2) && (21)\\ \sigma[X_1] &= \sigma_1\cdot\sqrt{1-\rho_{12}^2}\,. && (22) \end{aligned}$$

And, by symmetry, it holds

$$X_2|_{x_1} \sim \mathcal{N}\left(\mu_2 + \rho_{12}\,\frac{\sigma_2}{\sigma_1}\,(x_1-\mu_1)\,,\,\sigma_2\sqrt{1-\rho_{12}^2}\right)\,. \tag{23}$$

Mnemonic rules to remember Eqs. (20) and (21) are

- the shift of the expected value depends linearly on the correlation coefficient as well on the difference between the value of the conditionand (x_2) and its expected value (μ_2); the ratio σ_1/σ_2 can be seen as a minimal dimensional factor in order to get a quantity that has the same dimensions of μ_1 (remember that X_1 and X_2 have in general different physical dimensions);
- the variance is reduced by a factor which depends on correlation coefficient, but not on its sign. In particular it goes to zero if $|\rho_{12}| \to 1$, limit in which the two quantities become linear dependent, while it does not change if $\rho_{12} \to 0$, since the two variables become independent and they cannot effect each other.

An example of a bivariate distribution (from [4], with x_1 and x_2 indicated as customary with x and y) is given in Fig. 5, which shows also the marginals and some conditionals.

2.5.1 Evaluation of a conditional from a given bivariate normal

As an exercise, lets prove (19), with the purpose of show some useful tricks to simplify the calculations. If we take literally the rule to evaluate $f(x_1\,x_2\,|\,I)$ knowing that $f(x_1, x_2\,|\,I)$ is given by (7) we need to calculate

$$f(x_1\,|\,x_2,\,I) \;=\; \frac{f(x_1, x_2\,|\,I)}{f(x_2\,|\,I)}\,. \tag{24}$$

The trick is to make the calculations neglecting all irrelevant multiplicative factors, starting from the denominator $f(x_2\,|\,I)$, which is a number given $X_2 = x_2$ (whatever its value might be!).

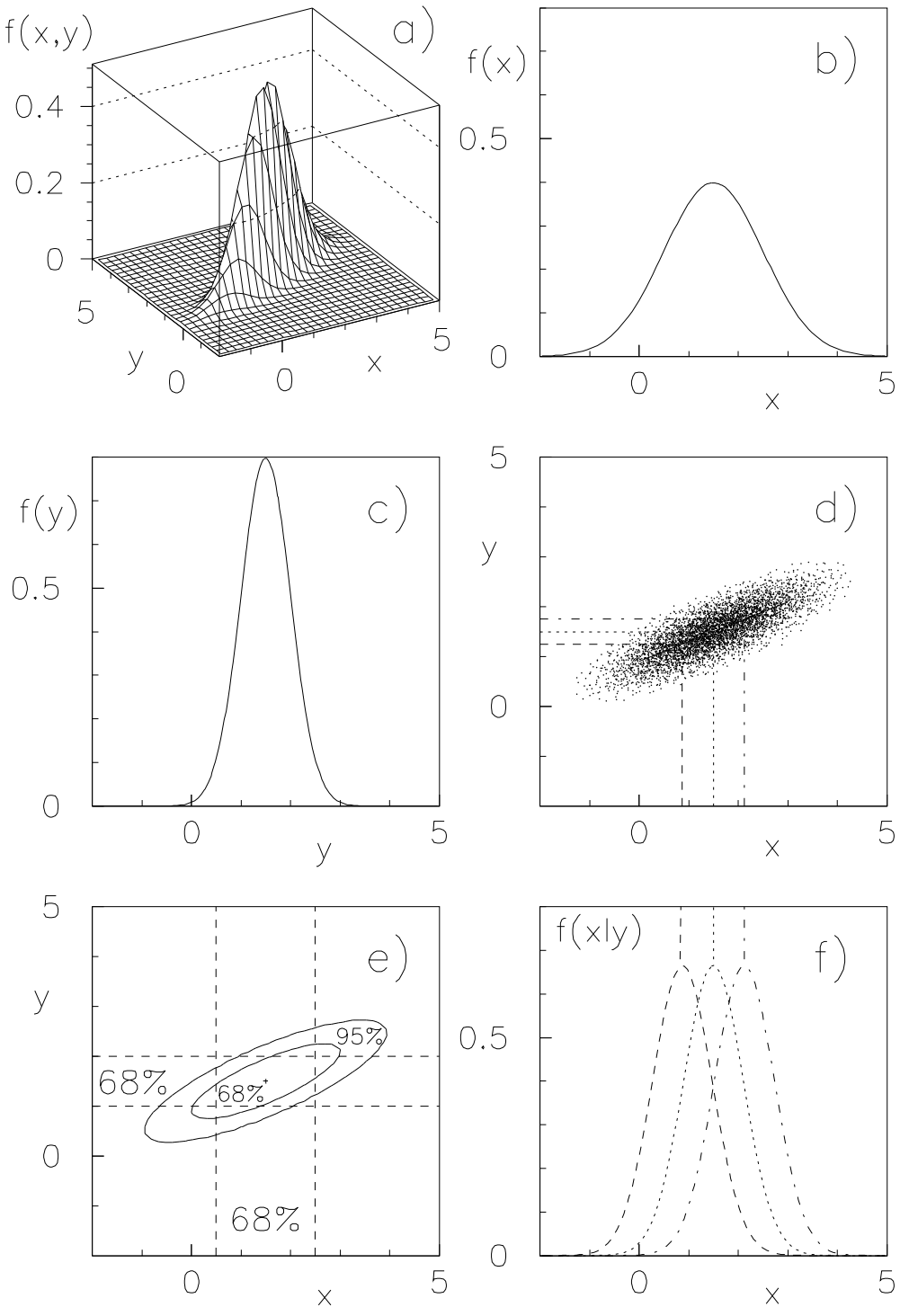


Figure 5: Example of bivariate normal distribution.

Here are the details (note that additive terms in the exponential are factors in the function of interest!):[9]

$$
\begin{aligned}
f(x_1 \,|\, x_2,\, I) \;&\propto\; f(x_1, x_2 \,|\, I) \\
&\propto\; \exp\left\{-\frac{1}{2\,(1-\rho_{12}^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} - 2\,\rho_{12}\,\frac{(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\,\sigma_2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right]\right\} \\
&\propto\; \exp\left\{-\frac{1}{2\,(1-\rho_{12}^2)\,\sigma_1^2}\left[(x_1-\mu_1)^2 - 2\,\rho_{12}\,\frac{\sigma_1}{\sigma_2}(x_1-\mu_1)(x_2-\mu_2)\right]\right\} \\
&\propto\; \exp\left\{-\frac{1}{2\,(1-\rho_{12}^2)\,\sigma_1^2}\left[x_1^2 - 2\,\mu_1 x_1 + \mu_1^2 - 2\,\rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(x_2-\mu_2)\,x_1\right]\right\}, \\
&\propto\; \exp\left\{-\frac{1}{2\,(1-\rho_{12}^2)\,\sigma_1^2}\left[x_1^2 - 2\,x_1\,[\mu_1 + \rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(x_2-\mu_2)]\right]\right\} \\
&\propto\; \exp\left\{-\frac{1}{2\,(1-\rho_{12}^2)\,\sigma_1^2}\left[x_1^2 - 2\,x_1\,[\mu_1 + \rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(x_2-\mu_2)] + [\mu_1 + \rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(x_2-\mu_2)]^2\right]\right\} \\
&\propto\; \exp\left\{-\frac{1}{2\,(1-\rho_{12}^2)\,\sigma_1^2}\left(x_1 - [\mu_1 + \rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(x_2-\mu_2)]\right)^2\right\} \qquad (25)
\end{aligned}
$$

in which we recognize a Gaussian with expected value $\mu_1 + \rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(x_2-\mu_2)$ and standard deviation $\sigma_1\sqrt{1-\rho_{12}^2}$ (and therefore the normalization factor can be obtained without any calculation).

2.6 Linear combinations

Linear transformations of variables are important because there are several practical problems to which they apply. There are also other cases in which the transformation is not rigorously linear, but it can be still linearized in the region of interest, where the probability mass is concentrated. There are well known theorems that relate expected values and covariance matrix of the *input quantities* to expected values and covariance matrix of the *output quantities*. The most famous case is when a single output quantity Y depend on several *independent variables* $\boldsymbol{X}$. So, given

$$
Y \;=\; \sum_i c_i\,X_i\,, \qquad (26)
$$

[9] Essentially the trick consists in observing that if we have a pdf proportional to $\exp\left[-h^2\,(x^2+\alpha\,x)\right]$, then it is also proportional to

$$
\exp\left[-h^2\,\left(x^2 + 2\,\frac{\alpha}{2}\,x + \left(\frac{\alpha}{2}\right)^2\right)\right] = \exp\left[-h^2\,\left(x - \left(-\frac{\alpha}{2}\right)\right)^2\right],
$$

that is a Gaussian with $\mu = -\alpha/2$ and $\sigma^2 = 1/(2\,h^2)$.

there is a relation which always holds, no matter if the X_i are independent or not and whichever are the pdf's which describe them:

$$\mathrm{E}[Y] \;=\; \sum_i c_i \,\mathrm{E}[X_i]\,. \tag{27}$$

In the special case that the X_i are also **independent**, we have

$$\mathrm{Var}[Y] \;=\; \sum_i c_i^2 \,\mathrm{Var}[X_i]. \tag{28}$$

Instead it is not always simple to calculate the pdf of Y in the most general case. There are however two remarkable cases, which we assume known and just recall them here, in which is Y is normally distributed:

1. **linear combinations of normally distributed variables** are still normal;
2. the **Central Limit Theorem** states that if we have 'many' **independent variables** (the theorem says "for n that goes to infinity"! Some practice is then needed to judge when it is large enough.) their linear combination is normally distributed with variance equal to $\sum_i c_i^2 \,\mathrm{Var}[X_i]$ if none of the non-normal components dominates the overall variance, i.e. if $c_j^2 \,\mathrm{Var}[X_j] \ll \sum_i c_i^2 \,\mathrm{Var}[X_i]$, where j denotes any of those non-normal components.

Since in this paper we only stick to normal distributed variables, the only task will be to evaluate the covariance matrix of the set of variables of interest, depending on the problem.

The general transformation from n input variables to m output variable is given by[10]

$$Y_i \;=\; c_{ij} X_j\,, \tag{29}$$

or, in a compact form that use the *transformation matrix* $\boldsymbol{C}$, whose elements are the c_{ij},

$$\boldsymbol{Y} \;=\; \boldsymbol{C}\,\boldsymbol{X}\,. \tag{30}$$

Expected value and covariance matrix of the output quantities are given by

$$\mathrm{E}[\boldsymbol{Y}] \;=\; \boldsymbol{C}\,\mathrm{E}[\boldsymbol{X}] \tag{31}$$

$$\boldsymbol{V}_Y \;=\; \boldsymbol{C}\,V_X\,\boldsymbol{C}^T \tag{32}$$

For example, if $\boldsymbol{\mu}_X = (2, -3)$, with $\sigma_{X_1} = 0.2$, $\sigma_{X_2} = 0.5$ and $\rho_{X_{12}} = -0.8$, and the transformation rule is given by

$$Y_1 \;=\; X_1 + 2\,X_2 \tag{33}$$

$$Y_2 \;=\; -X_1 + X_2\,, \tag{34}$$

[10] We neglect a possible extra constant term in the linear combination because this plays no role in the uncertainty.

i.e.

$$\boldsymbol{C} \;=\; \begin{pmatrix} 1 & 2 \\ -1 & 1 \end{pmatrix} \tag{35}$$

we get in R: [11]

```
> mu.X <- c(2, -3)
> s.X <- c(0.2, 0.5)
> rho.X   <- -0.8
> V.X  <- outer(s.X, s.X)
> V.X[1,2] <- V.X[2,1] <- V.X[1,2]*rho.X
> V.X
      [,1]  [,2]
[1,]  0.04 -0.08
[2,] -0.08  0.25
> ( cor.X <- V.X / outer(s.X,s.X) )
     [,1] [,2]
[1,]  1.0 -0.8
[2,] -0.8  1.0
> ( C <- rbind( c(1,2), c(-1,1) ) )
     [,1] [,2]
[1,]    1    2
[2,]   -1    1
> ( mu.Y <- as.vector( C %*% mu.X ) )
[1] -4 -5
> ( V.Y <- C %*% V.X %*% t(C) )
     [,1] [,2]
[1,] 0.72 0.54
[2,] 0.54 0.45
> ( s.Y   <- sqrt(diag(V.Y)) )
[1] 0.8485281 0.6708204
> ( rho.Y <- V.Y[1,2] / prod(s.Y) )
[1] 0.9486833
> ( cor.Y <- V.Y / outer(s.Y,s.Y) )
          [,1]      [,2]
[1,] 1.0000000 0.9486833
[2,] 0.9486833 1.0000000
```

Let us get a visual representation of the probability distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$ using the random generator provided by the package mnormt (see result in Fig. 6):

[11] The function `outer()` produces by default a matrix which is by default is the *outer* product of two vectors, i.e. $\boldsymbol{v}_1\,\boldsymbol{v}_2^T$. But it has a third parameter `FUN` which which it is possible to evaluate different function on the 'grid' defined by the Cartesian product of the two vector. Try for example

```
> outer(1:3, 1:3, '+')
> outer(1:3, 1:3, function(x,y) x + y^2))
> round( outer(0:10, 0:10, function(x,y) sin(x)*cos(y)), 2 )
```

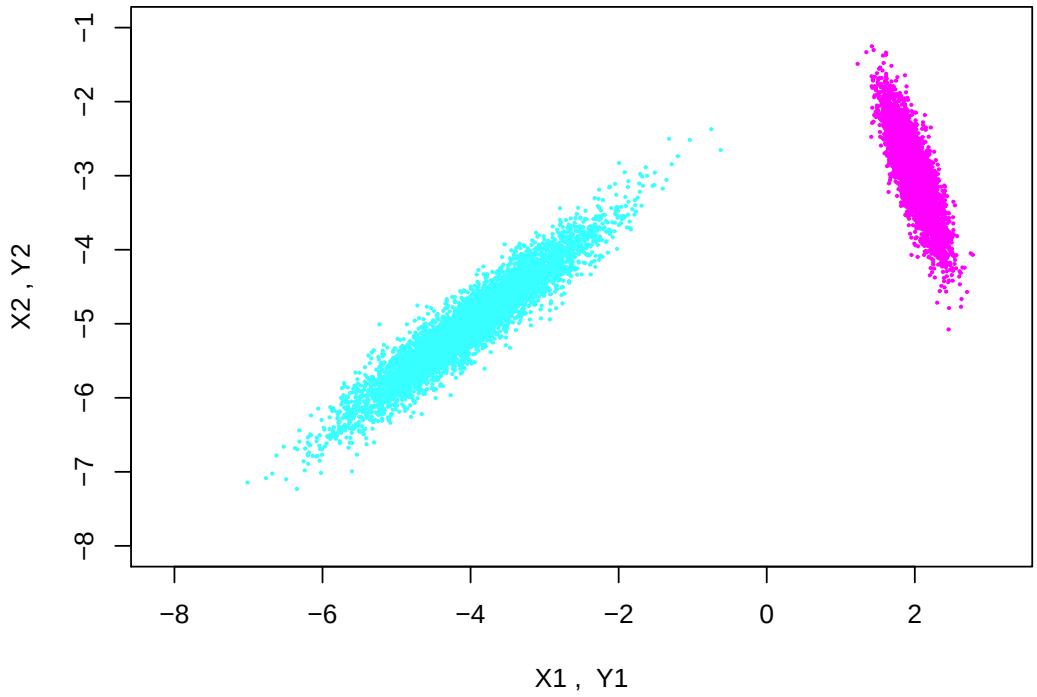


Figure 6: Monte Carlo sampling of two multivariate normal distributions (see text).

```
> n=5000; r.X <- rmnorm(n, mu.X, V.X); r.Y <- rmnorm(n, mu.Y, V.Y)
> plot(r.X, col='magenta', xlim=c(-7,2), ylim=c(-8,-1), cex=0.2,
+  asp=1, xlab='X1 ,  Y1',  ylab='X2 , Y2')
> points(r.Y, col='cyan', cex=0.2)
```

2.7 Conditional distributions in many dimensions

Instead, a less known rule is that that gives the covariance matrix of a conditional distribution with a number of variables above two. For example we might have 5 variables $X_1, X_2, \ldots X_5$ and could be interested in the expected values and the covariance matrix of (X_1, X_4 X_5), given (X_2, X_3). Problems of this kind might look a mere mathematical curiosity, but they are indeed important to understand how we learn from data and we make probabilistic predictions using probability theory.

Compact formulae to solve this problems are due to Morris Eaton [3]. If we partition $\boldsymbol{\mu}$ and $\boldsymbol{V}$ into the the subsets of variable on which we want to condition and the other ones,

i.e.

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \tag{36}$$

$$V = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix} \tag{37}$$

the result is

$$\mathrm{E}\left[\boldsymbol{X}_1|_{\boldsymbol{X}_2=\boldsymbol{a}}\right] = \boldsymbol{\mu}_1 + \boldsymbol{V}_{12}\,\boldsymbol{V}_{22}^{-1}\,(\boldsymbol{a}-\boldsymbol{\mu}_2) \tag{38}$$

$$\boldsymbol{V}\left[\boldsymbol{X}_1|_{\boldsymbol{X}_2}\right] = \boldsymbol{V}_{11} - \boldsymbol{V}_{12}\,\boldsymbol{V}_{22}^{-1}\,\boldsymbol{V}_{21} \tag{39}$$

(And analogous formulae for $\mathrm{E}\left[\boldsymbol{X}_2|_{\boldsymbol{X}_1=\boldsymbol{b}}\right]$ and $\mathrm{Var}\left[\boldsymbol{X}_2|_{\boldsymbol{X}_1=\boldsymbol{b}}\right]$.)

In the case of a bivariate distributions we recover easily Eqs. (20)-(21), as it follows.

Expected value: $\boldsymbol{V}_{12}$ is the off-diagonal term $\rho_{12}\sigma_1\sigma_2$, while $\boldsymbol{V}_{22}$ is equal to σ_2^2. Eq. (38) becomes then

$$\begin{aligned} \mathrm{E}[X_1|_{X_2}] &= \mu_1 + \rho_{12}\,\sigma_1\,\sigma_2\,\frac{1}{\sigma_2^2}\,(a-\mu_2) \\ &= \mu_1 + \rho_{12}\,\frac{\sigma_1}{\sigma_2}\,(a-\mu_2) \end{aligned} \tag{40}$$

Variance: The remaining two terms of interest are also very simple: $\boldsymbol{V}_{11}$ is σ_1^2, while $\boldsymbol{V}_{21}$ is equal to $\boldsymbol{V}_{12}$. It follows

$$\begin{aligned} \mathrm{Var}[X_1|_{X_2}] &= \sigma_1^2 - \rho_{12}\,\sigma_1\,\sigma_2\,\frac{1}{\sigma_2^2}\,\rho_{12}\,\sigma_1\,\sigma_2 \\ &= \sigma_1^2 - \rho_{12}^2\,\sigma_1^2 \\ &= \sigma_1^2\,(1-\rho_{12}^2). \end{aligned} \tag{41}$$

Note that, while the conditioned expected value depends on the conditionand vector $\boldsymbol{a}$, the conditioned variance does not.

3 R implementation of the rule to condition multivariate normal distributions

At this point, having set up all our tools, here is the R function which implements the above formulae:

```
norm.mult.cond <- function(mu, V, x.c, full=TRUE) {
  out <- NULL
  # Checks:
  n <- length(mu)
  # 1) dimensions of V
  if ( sum(dim(V) != n) ) {
    cat( sprintf("dimensions of V incompatible with length of mu\n") )
    return(out)
  }
  # 2) V must be positively defined (no negative eigenvalues )
  if( sum( eigen(V)$values <= 0 ) >  0) {
    cat( sprintf("V is not positively defined\n") )
    return(out)
  }
  # number of conditionand variables
  nc <- length(x.c[!is.na(x.c)])
  # peculiar/anomalous cases
  if( (length(x.c) > n) | (nc > n) ) {
    cat( sprintf("x.c has more elements than mu\n") )
    return(out)
  } else if (nc == 0) {  # No condition
    out$mu <- mu
    out$V  <- V
    return(out)
  } else if(nc == n) {
    out$mu <- x.c    # exact values
    out$V  <- NULL   # covariance matrix is meaningless
    return(out)
  }

  # Apply Eaton's formulae (-> Wiki)
  v.c <- which(!is.na(x.c))  # conditioning variables
  v   <- which(is.na(x.c))   # variables of interest
  V11 <- V[v, v]      # Sigma_11 in Wiki
  V22 <- V[v.c, v.c]  # Sigma_22  "  "
  V12 <- V[v, v.c]    # Sigma_12  "  "
  V21 <- V[v.c, v]    # Sigma_21  "  "
  mu.cond <- mu[v] + V12 %*% solve(V22) %*%  (x.c[!is.na(x.c)] - mu[v.c])
  V.cond  <- V11 - V12 %*% solve(V22) %*% V21
  if(!full) { # returns only interesting part
    out$mu <- as.vector(mu.cond)
    out$V  <- V.cond
  } else {    # returns all (better to understand!!)
    mu1 <- mu
    V1 <- V
    mu1[v]   <- mu.cond
    mu1[v.c] <- x.c[!is.na(x.c)]
```

```
    V1[v, v] <- V.cond
    V1[v.c, v.c] <- 0
    V1[v, v.c]   <- 0
    V1[v.c, v]   <- 0
    out$mu <- as.vector(mu1)
    out$V  <- V1
  }
  return(out)
}
```

(The conditionand vector `x.c` has to contain numbers in the positions corresponding to the variables on which we want to condition, and `NA`, i.e. 'not available', or 'unknown', in the others, as we shall see in the examples.)

Let us try with a simple case of two normal quantities $\boldsymbol{\mu}_X = (2, -3)$ of section 2.6. The question is how our uncertainty on μ_{X_1} change if *we assume* $\mu_{X_2} = -2$:

```
> ( V.X.cond <- norm.mult.cond(mu.X, V.X, c(NA, -2)) )
$mu
[1]  1.68 -2.00

$V
       [,1] [,2]
[1,] 0.0144    0
[2,] 0.0000    0

> sqrt(diag(V.X.cond$V))
[1] 0.12 0.00
```

The effect of the conditionis to shift the expected value of μ_{X_1} from 2 to 1.68 and to squeeze its *standard uncertainty* to 0.12. If we provide our result in the conventional form "expected value $\pm$ standard uncertainty", the assumption (or 'knowledge') $X_2 = -2$ updates our 'knowlwdge' about X_1 from '2.00 ± 0.20' to '1.67 ± 0.12'.

4 The 'simplest experiment'

Let us go back to the first diagram of Fig. 1, that we repeat here for convenience:

It describes the situation in which we have the physical quantity X_1, that is a parameter of our physical model of reality, and the reading on an instrument, X_2, *caused* by X_1.

The instrument has been well calibrate, such to give X_2 around X_1, but it is not perfect, as usual. In other words, even if we new X_1 we were not sure about the value we would read. For simplicity, let us model this uncertainty by a normal distribution, i.e.

$$X_2|_{X_1} \quad \sim \quad \mathcal{N}(X_1, \sigma_{2|1})\,. \tag{42}$$

But we usually do not know X_1, and therefore we are even more uncertain about what we shall read on the instrument. In fact we are dealing with a joint distribution describing the joint uncertainty about the two quantities, that is

$$f(x_1,\, x_2\,|\,I) \quad = \quad f(x_2\,|\,x_1,\, I)\cdot f(x_1\,|\,I)\,. \tag{43}$$

Our knowledge about X_2 will be given, instead, by $f(x_2\,|\,I) = \int_{\{x_1\}} f(x_1,\, x_2\,|\,I)\,dx_1$, a distribution characterized by $\mathrm{Var}[X_2] \neq \mathrm{Var}[X_2|_{X_1}]$.

It is convenient to model our uncertainty about X_1 with a normal distribution, with a standard deviation σ_1 much larger than $\sigma_{2|1}$ – if we make a measurement we want to gain knowledge about that quantity! – and centered around the values we roughly expect.[12]

In order to simplify the calculations, in the exercise that follows let us assume that X_1 is centered around zero. We shall see later how to get rid of this limitation.

The joint distribution $f(x_1,\, x_2\,|\,I)$ is then given by

$$f(x_1,\, x_2\,|\,I) \quad = \quad \frac{1}{\sqrt{2\,\pi}\,\sigma_{2|1}}\,\exp\left[-\frac{(x_2-x_1)^2}{2\,\sigma_{2|1}^2}\right]\times\frac{1}{\sqrt{2\,\pi}\,\sigma_1}\,\exp\left[-\frac{x_1^2}{2\,\sigma_1^2}\right] \tag{44}$$

As an exercise, let us see how to evaluate $f(x_1,\, x_2\,|\,I)$. The trick, already applied before, is to manipulate the terms in the exponent in order to recover some well known pattern.

[12] For extensive discussions about modelling prior knowledge of physical quantities see Ref. [4] and references therein. As a practical example, think at the width of the table at which a sit in the very moment you read these lines (or any other object), and about the reading on a ruler when you try to measure it.

Here are the details, starting from (44) rewritten dropping all irrelevant factors:

$$
\begin{aligned}
f(x_1,\, x_2 \,|\, I) &\propto \exp\left[-\frac{(x_2-x_1)^2}{2\,\sigma_{2|1}^2}-\frac{x_1^2}{2\,\sigma_1^2}\right] && (45)\\
&\propto \exp\left[-\frac{1}{2}\left(\frac{x_2^2-2\,x_1x_2+x_1^2}{\sigma_{2|1}^2}+\frac{x_1^2}{\sigma_1^2}\right)\right] && (46)\\
&\propto \exp\left[-\frac{1}{2}\left(\frac{x_2^2}{\sigma_{2|1}^2}-\frac{2\,x_1x_2}{\sigma_{2|1}^2}+x_1^2\cdot\left(\frac{1}{\sigma_{2|1}^2}+\frac{1}{\sigma_1^2}\right)\right)\right] && (47)\\
&\propto \exp\left[-\frac{1}{2}\left(\frac{x_2^2}{\sigma_{2|1}^2}-\frac{2\,x_1x_2}{\sigma_{2|1}^2}+x_1^2\cdot\frac{\sigma_{2|1}^2+\sigma_1^2}{\sigma_{2|1}^2\cdot\sigma_1^2}\right)\right] && (48)\\
&\propto \exp\left[-\frac{1}{2}\,\frac{\sigma_{2|1}^2+\sigma_1^2}{\sigma_{2|1}^2}\,\left(\frac{x_2^2}{\sigma_{2|1}^2+\sigma_1^2}-\frac{2\,x_1x_2}{\sigma_{2|1}^2+\sigma_1^2}+\frac{x_1^2}{\sigma_1^2}\right)\right] && (49)\\
&\propto \exp\left[-\frac{1}{2}\,\frac{1}{\frac{\sigma_{2|1}^2}{\sigma_{2|1}^2+\sigma_1^2}}\,\left(\frac{x_2^2}{\sigma_{2|1}^2+\sigma_1^2}-\frac{2\,x_1x_2}{\sigma_{2|1}^2+\sigma_1^2}+\frac{x_1^2}{\sigma_1^2}\right)\right] && (50)
\end{aligned}
$$

In this expression we recognize a bivariate distribution centered around $(0,0)$, provided we interpret

$$
\begin{aligned}
\sigma_{2|1}^2+\sigma_1^2 &= \sigma_2^2 && (51)\\
\frac{\sigma_{2|1}^2}{\sigma_{2|1}^2+\sigma_1^2} &= 1-\rho_{12}^2\,, && (52)
\end{aligned}
$$

and after having checked the consistency of the terms multiplying $x_1\,x_2$. Indeed we have

$$
\begin{aligned}
\rho_{12}^2 &= 1-\frac{\sigma_{2|1}^2}{\sigma_{2|1}^2+\sigma_1^2}=\frac{\sigma_1^2}{\sigma_{2|1}^2+\sigma_1^2} && (53)\\
\rho_{12} &= \frac{\sigma_1}{\sqrt{\sigma_{2|1}^2+\sigma_1^2}}=\frac{\sigma_1}{\sigma_2} && (54)
\end{aligned}
$$

and then the second term within parenthesis can be rewritten as

$$
\frac{2\,x_1x_2}{\sigma_{2|1}^2+\sigma_1^2} = \frac{2\,x_1x_2}{\sigma_2\cdot\sigma_2}=\frac{2\,\rho_{12}\,x_1x_2}{\sigma_1\cdot\sigma_2}. \qquad (55)
$$

Then

$$
f(x_1,\, x_2 \,|\, I) \propto \exp\left[-\frac{1}{2\,(1-\rho_{12}^2)}\left(\frac{x_1^2}{\sigma_1^2}-\frac{2\,\rho_{12}\,x_1x_2}{\sigma_1\cdot\sigma_2}+\frac{x_2^2}{\sigma_2^2}\right)\right] \qquad (56)
$$

is definitively a bivariate normal distribution with

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{57}$$

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_{2|1}^2 \end{pmatrix} \tag{58}$$

As a cross check, let us evaluate expected value and variance of X_2 if we assume a certain value of X_1, for example $X_1 = x_1$:

$$\mathrm{E}[X_2|_{X_1=x_1}] = 0 + \frac{\sigma_1^2}{\sigma_1^2} \cdot (x_1 - 0) = x_1 \tag{59}$$

$$\mathrm{Var}[X_2|_{X_1=x_1}] = \sigma_1^2 + \sigma_{2|1}^2 - \frac{\sigma_1^2}{\sigma_1^2}\,\sigma_1^2 = \sigma_{2|1}^2\,, \tag{60}$$

as it should be: provided we know the value of X_1 our expectation of X_2 is around its value, with standard uncertainty $\sigma_{2|1}$.

More interesting is the other way around, that is indeed the purpose of the experiment: how our knowledge about X_1 is modified by $X_2 = x_2$:

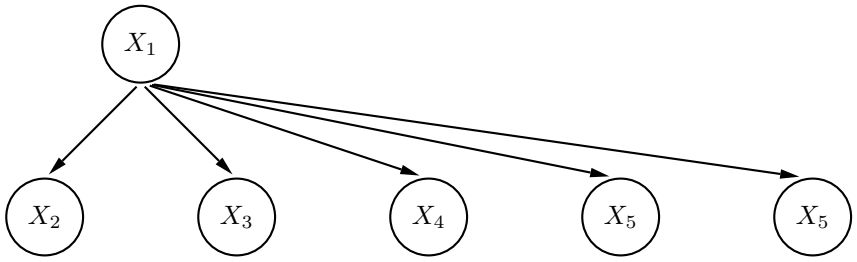
$$\mathrm{E}[X_1|_{X_2=x_2}] = 0 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_{1|2}^2} \cdot (x_2 - 0) = x_2 \cdot \frac{1}{1 + \sigma_{1|2}^2/\sigma_1^2} \tag{61}$$

$$\mathrm{Var}[X_1|_{X_2=x_2}] = \sigma_1^2 - \frac{\sigma_1^2}{\sigma_1^2 + \sigma_{1|2}^2}\,\sigma_1^2 = \sigma_{1|2}^2 \cdot \frac{1}{1 + \sigma_{1|2}^2/\sigma_1^2}\,, \tag{62}$$

Contrary to the first case, this second result is initially not very intuitive: the expected value of X_1 is not exactly equal to the 'observed' value x_2, unless σ_1, that models our *prior standard uncertainty* about X_1, is much larger than the experimental resolution $\sigma_{2|1}$. Similarly, the *final standard uncertainty* is in general a bit smaller than $\sigma_{2|1}$, unless, again, $\sigma_{1|2}/\sigma_1 \ll 1$. Although initially surprising, these result are in qualitative agreement with the good sense of experienced physicists [4].

5 Several independent measurements on the same physics quantity

The next step is to see what happens when we are in the conditions to make several *independent measurements* on the same quantity X_1, possibly with different instruments, each one characterized by a conditional standard uncertainty $\sigma_{i|1}$ and perfectly calibrated, that is $\mathrm{E}[X_i|_{X_1=x_1}] = x_1$. The situation can be illustrated with the diagram at the center of Fig. 1, reported here for convenience, extendend to other observations:



We have learned that if we are able to build up the covariance matrix of the joint distribution $f(x_1, x_2, x_3 \,|\, I)$ the problem is readily solved, at least in the normal approximations we are using throughout the paper.

In principle we should repeat the previous exercise to evaluate

$$\begin{aligned} f(x_1, x_2, x_3 \,|\, I) &= f(x_1 \,|\, I) \cdot f(x_2 \,|\, x_1,\, I) \cdot f(x_3 \,|\, x_1, x_2\, I) && (63) \\ &= f(x_1 \,|\, I) \cdot f(x_2 \,|\, x_1,\, I) \cdot f(x_3 \,|\, x_1,\, I)\,, && (64) \end{aligned}$$

where in the last step we have made explicit that $f(x_3 \,|\, x_1\, I)$ does not depend on X_2, once X_1 is known (but this does not implies that X_2 and X_3 are independent, as we shall see later! They are simply *conditionally independent*, i.e. independent under the condition that X_1 has a given value.)

In reality we do not need to go through a similar derivation, that indeed *was just an exercise*. The easy solution arises, going back to the previous case, noting that X_2 could be considered just the sum of X_1 and $X_2|_{X_1}$. Therefore we can just use the rules of linear transformations of normal multivariates, which can easily extended to any number of observations. Here is the transformation rule for three variables

$$\begin{aligned} Y_1 &= X_1 && (65) \\ Y_2 &= X_1 + X_2|_{X_1} && (66) \\ Y_3 &= X_1 + X_3|_{X_1} && (67) \end{aligned}$$

from which the calculation of the covariance matrix is straightforward:

- the i-th element diagonal is given by the variance of Y_i, that is σ_1^2, $(\sigma_1^2 + \sigma_{2|1}^2)$, and so on;
- the off-diagonal elements are all equal to σ_1^2, because the only element in common in all linear combinations is X_1.

Hence here is the covariance matrix of interest:

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_{2|1}^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 & \sigma_1^2 + \sigma_{3|1}^2 \end{pmatrix} \qquad (68)$$

5.1 Getting some insights with numerical examples

At this point, instead of trying to get analytic formulae for all conditional probabilities of interest, we prefer to use the properties of the multivariate normal distribution implemented in the function `norm.mult.cond()` seen before. And, since the game is now automatic, we enlarge our space to 6 variables, X_1 for the 'true value' and X_2 -X_6 for four possible readings. Although it is not any longer needed, we still set out prior central value about X_1 around 0, which is equivalent to set to 0 all expected values. Here is the R code, with some comments:

```
> n=6; muX1=0; sigmaX1=10                # set size and initial uncertainty on X1
> mu <- rep(muX1, n)                     # set expected values (all equal!)
> ( sigma <- c(sigmaX1, rep(1,n-1)) )    # standard deviations
[1] 10  1  1  1  1  1
> V <- matrix(rep(sigma[1]^2, n*n), c(n,n))
> diag(V)[2:n] <- diag(V)[2:n] + sigma[2:n]^2
> V                                                  # covariance matrix
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]  100  100  100  100  100  100
[2,]  100  101  100  100  100  100
[3,]  100  100  101  100  100  100
[4,]  100  100  100  101  100  100
[5,]  100  100  100  100  101  100
[6,]  100  100  100  100  100  101
> (su <- sqrt(diag(V)))                              # standard deviations
[1] 10.00000 10.04988 10.04988 10.04988 10.04988 10.04988
> V/outer(su,su)                                     # correlation matrix
          [,1]      [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 1.0000000 0.9950372 0.9950372 0.9950372 0.9950372 0.9950372
[2,] 0.9950372 1.0000000 0.9900990 0.9900990 0.9900990 0.9900990
[3,] 0.9950372 0.9900990 1.0000000 0.9900990 0.9900990 0.9900990
[4,] 0.9950372 0.9900990 0.9900990 1.0000000 0.9900990 0.9900990
[5,] 0.9950372 0.9900990 0.9900990 0.9900990 1.0000000 0.9900990
[6,] 0.9950372 0.9900990 0.9900990 0.9900990 0.9900990 1.0000000
```

As we can see, all variables are correlated! The reason is very simple: any precise information we get about one of them changes the pdf of all others. In physics terms, a reading on a instrument changes our opinion about the true value of the quantity of interest as well as of all other readings we have not yet done (or we not aware of their values – in probability theory what matters is not time ordering, but ignorance).

Let us now see what happens if we condition on a **precise value of** X_1, for example $X_1 = 2$:

```
> ( mu.c <- c(2, rep(NA, n-1)) )                  # conditionand
[1]  2 NA NA NA NA NA
> ( out<- norm.mult.cond(mu, V, mu.c) )           # resulting multivariate
$mu
[1] 2 2 2 2 2 2
```

```
$V
     [,1] [,2] [,3] [,4] [,5] [,6]
[1,]    0    0    0    0    0    0
[2,]    0    1    0    0    0    0
[3,]    0    0    1    0    0    0
[4,]    0    0    0    1    0    0
[5,]    0    0    0    0    1    0
[6,]    0    0    0    0    0    1
```

As we see, the expected values are all equal, X_1 is not longer uncertain, and all other variables become *independent*,[13] more precisely "conditional independent"

Let's now see what happens if we condition on the **observation** $\underline{X_2 = 2}$:

```
> ( mu.c <- c(NA, 2, rep(NA, n-2)) )
[1] NA  2 NA NA NA NA
> ( out<- norm.mult.cond(mu, V, mu.c) )
$mu
[1] 1.980198 2.000000 1.980198 1.980198 1.980198 1.980198

$V
          [,1] [,2]     [,3]     [,4]     [,5]     [,6]
[1,] 0.990099    0 0.990099 0.990099 0.990099 0.990099
[2,] 0.000000    0 0.000000 0.000000 0.000000 0.000000
[3,] 0.990099    0 1.990099 0.990099 0.990099 0.990099
[4,] 0.990099    0 0.990099 1.990099 0.990099 0.990099
[5,] 0.990099    0 0.990099 0.990099 1.990099 0.990099
[6,] 0.990099    0 0.990099 0.990099 0.990099 1.990099

> ( out.s <- sqrt(diag(out$V)) )                 # standard deviations
[1] 0.9950372 0.0000000 1.4107087 1.4107087 1.4107087 1.4107087
> out$V / outer(out.s, out.s)                    # correlation matrix (besides NaN)
          [,1] [,2]      [,3]      [,4]      [,5]      [,6]
[1,] 1.0000000  NaN 0.7053456 0.7053456 0.7053456 0.7053456
[2,]       NaN  NaN       NaN       NaN       NaN       NaN
[3,] 0.7053456  NaN 1.0000000 0.4975124 0.4975124 0.4975124
[4,] 0.7053456  NaN 0.4975124 1.0000000 0.4975124 0.4975124
[5,] 0.7053456  NaN 0.4975124 0.4975124 1.0000000 0.4975124
[6,] 0.7053456  NaN 0.4975124 0.4975124 0.4975124 1.0000000
```

The effect of the 'measurement' has changed all our expectations, all 'practically equal' to the observed value of 2, but the uncertainties about the possible 'future measurements' are $\sqrt{2}$ larger than those of X_1 (Fig. 7). The reason is that X_2 and X_3 and all other possible readings 'communicate via' X_1: their uncertanty is than the combination (quadratic combination!) of that assigned to X_1 and the that of the readings if we new exactly X_1 (that is $\sigma_{i|1}$).

Let us see if we add **another observation**, e.g. $\underline{X_3 = 1}$:

[13] In general independence implies null covariance. For normal multivariate is true also the opposite.

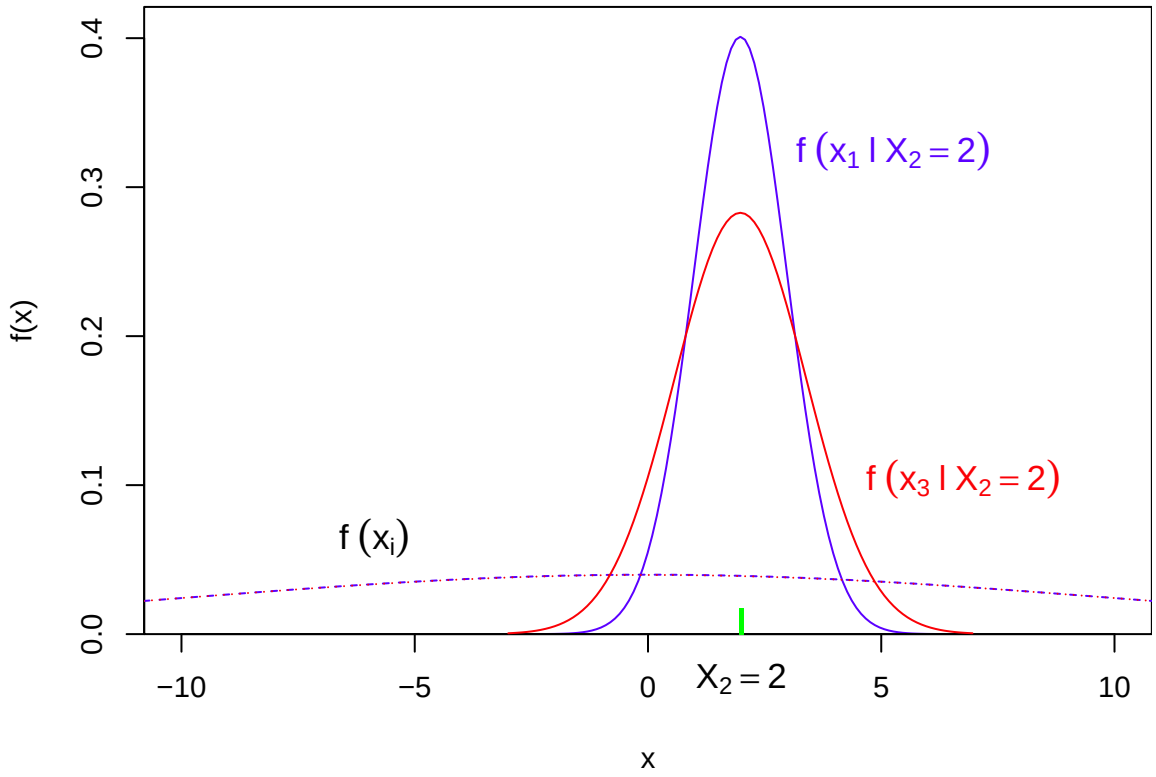


Figure 7: Normal distributions describing our uncertainty about X_1 and X_3 before (dashed line) and after (solid line) the observation $X_2 = 2$ (see text).

```
> mu.c <- c(NA, 2, 1, NA, NA, NA)
> ( out<- norm.mult.cond(mu, V, mu.c) )
$mu
[1] 1.492537 2.000000 1.000000 1.492537 1.492537 1.492537

$V
          [,1] [,2] [,3]      [,4]      [,5]      [,6]
[1,] 0.4975124    0    0 0.4975124 0.4975124 0.4975124
[2,] 0.0000000    0    0 0.0000000 0.0000000 0.0000000
[3,] 0.0000000    0    0 0.0000000 0.0000000 0.0000000
[4,] 0.4975124    0    0 1.4975124 0.4975124 0.4975124
[5,] 0.4975124    0    0 0.4975124 1.4975124 0.4975124
[6,] 0.4975124    0    0 0.4975124 0.4975124 1.4975124

> ( out.s <- sqrt(diag(out$V)) )
[1] 0.7053456 0.0000000 0.0000000 1.2237289 1.2237289 1.2237289

> out$V / outer(out.s, out.s)
```

```
          [,1] [,2] [,3]      [,4]      [,5]      [,6]
[1,] 1.0000000  NaN  NaN 0.5763904 0.5763904 0.5763904
[2,]       NaN  NaN  NaN       NaN       NaN       NaN
[3,]       NaN  NaN  NaN       NaN       NaN       NaN
[4,] 0.5763904  NaN  NaN 1.0000000 0.3322259 0.3322259
[5,] 0.5763904  NaN  NaN 0.3322259 1.0000000 0.3322259
[6,] 0.5763904  NaN  NaN 0.3322259 0.3322259 1.0000000
```

As we can see, after the second observation the expected values are practicalli equal to 1.5, average between the two readings. The uncertainty about the true value has decreased by a factor 1.41, that is $\sqrt{2}$, while the uncertainties about the forecastings decrease only by a factor 1.15, going from 1.41 to 1.22. This latter number can be understood as $\sqrt{0.705^2 + 1^2} = 1.22$, as it will be justified in a while.

Let us see what happens if we suppose that X_1 is well known:

```
> mu.c <- c(3, 2, 1, NA, NA, NA)
> ( out<- norm.mult.cond(mu, V, mu.c) )
$mu
[1] 3 2 1 3 3 3

$V
     [,1] [,2] [,3]          [,4]          [,5]          [,6]
[1,]    0    0    0  0.000000e+00  0.000000e+00  0.000000e+00
[2,]    0    0    0  0.000000e+00  0.000000e+00  0.000000e+00
[3,]    0    0    0  0.000000e+00  0.000000e+00  0.000000e+00
[4,]    0    0    0  1.000000e+00 -2.302158e-12 -2.302158e-12
[5,]    0    0    0 -2.302158e-12  1.000000e+00 -2.302158e-12
[6,]    0    0    0 -2.302158e-12 -2.302158e-12  1.000000e+00
```

If X_1 is perfectly known the observations X_2 and X_3 are irrelevant, as it has to be.[14]

Finally, let us add a **third observation**, e.g. $\underline{X_4 = 0}$

```
> mu.c <- c(NA, 2, 1, 0, NA, NA)
> ( out<- norm.mult.cond(mu, V, mu.c) )
$mu
[1] 0.9966777 2.0000000 1.0000000 0.0000000 0.9966777 0.9966777

$V
          [,1] [,2] [,3] [,4]      [,5]      [,6]
[1,] 0.3322259    0    0    0 0.3322259 0.3322259
[2,] 0.0000000    0    0    0 0.0000000 0.0000000
[3,] 0.0000000    0    0    0 0.0000000 0.0000000
[4,] 0.0000000    0    0    0 0.0000000 0.0000000
[5,] 0.3322259    0    0    0 1.3322259 0.3322259
[6,] 0.3322259    0    0    0 0.3322259 1.3322259
```

[14] And if X_2 and X_3 are 'very far' from X_1? In this simple model we are using, there is little to do, because any observation from minus infinite to plus infinite is never incompatible with a any Gaussian. But we know by experience that something strange might be happened. It this case we need to put in mathematical form the model we have in mind.

```
> ( out.s <- sqrt(diag(out$V)) )
[1] 0.5763904 0.0000000 0.0000000 0.0000000 1.1542209 1.1542209
> out$V / outer(out.s, out.s)
          [,1] [,2] [,3] [,4]      [,5]      [,6]
[1,] 1.0000000  NaN  NaN  NaN 0.4993762 0.4993762
[2,]       NaN  NaN  NaN  NaN       NaN       NaN
[3,]       NaN  NaN  NaN  NaN       NaN       NaN
[4,]       NaN  NaN  NaN  NaN       NaN       NaN
[5,] 0.4993762  NaN  NaN  NaN 1.0000000 0.2493766
[6,] 0.4993762  NaN  NaN  NaN 0.2493766 1.0000000
```

As we can see, the value of X_1 is with very good approximation the average of the three observations, that is 1, with a the standard uncertainty decreasing with $1/\sqrt{n}$, passing from 1.00 to 0.71 to 0.58. This is because the three pieces of information play the same wight, since $\sigma_{i|1}$, related to the 'precision of the instrument', is the same in all cases and equal to 1.

As far as the prediction of future observations, obviously they must be centered around the value we think X_1, is at the best of our knowledge, a value which changes with the observations. As far as the uncertainty and correlation coefficient, they decrease as follows (starting from the very beginning, before any observation):

Standard uncertainty: 10.05, 1.41, 1.22, 1.15.

We can see that they are a quadratic combination of the uncertainty with which we know X_1 and that with which we expect the observation given a precise value of X_1. If we indicate the state of information at time t as $I(t)$, the rule is

$$\mathrm{Var}[X_i \,|\, I(t)] \;=\; \mathrm{Var}[X_1 \,|\, I(t)] + \sigma^2_{i|1}\,. \tag{69}$$

Asymptotically, when after many measurements the determination of X_1 is very accurate, it only remains $\sigma^2_{i|1}$, as it has to be.

Correlation coefficient: 0.990, 0.50, 0.33, 0.25.

It is initially very high because any precise value of each observation changes dramatically our expectation about the others. But then, when we have already made several observations a new one has only very little effect on our forecasting. Asymptotically, when we have made a very large number of observations and X_1 is very well 'determined', all future observations become essentially "conditionally independent".

5.2 Follows up

At this point the game can be continued with different options. One has only to re-build the initial covariance matrix and play changing the conditions.

An interesting exercise is to increase σ_1, for example to 100, i.e. 100 times large than the 'precision' of our instrument, to see how our conclusions change.

It could also interesting to see what happens if the different observations come from instruments having different precisions.

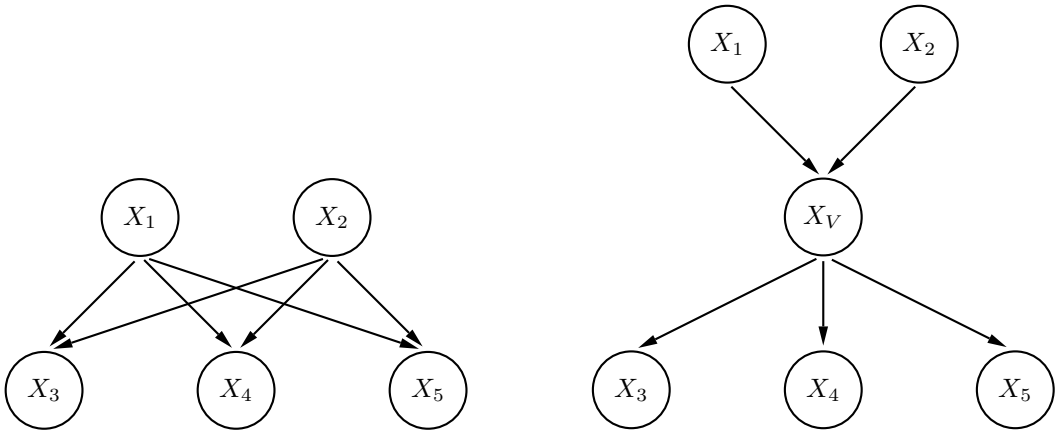
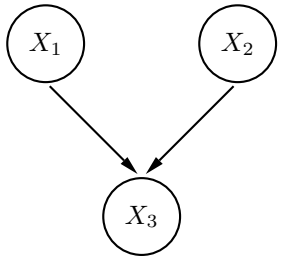


Figure 8: Basic models of joint probabilities

6 Adding a simple systematic effect ('offset')

Let us now move to the second diagrams, in which X_3 gets influenced by X_1 and X_2:



This can model the presence of a systematic effect, because we expect that the possible values of X_3 are caused by both X_1 and X_2, and it will be then influenced by how uncertain is the quantity X_2 that acts as a systematic. The simplest case of systematic effect is an additive one, of unknown value, but with expected value 0 (the instrument has been calibrated 'at the best'!) and a standard uncertainty σ_2. Needless to say, we also model this uncertainty with a normal distribution, with much simplification in the calculations (and also because this is often the case).

The model can be extended to several observations, as shown in the left diagram of Fig. 8. In the figure it is also shown a different interpretation of the effect of the systematic error, which is very close to the physicist intuition. The observations X_3, X_4 and X_5 are normally distributed around a kind of 'virtual state' X_V determined by the *unknown* true value X_1 and the *unknown* offset X_2. The transformation rule to build the initial covariance

matrix will be then

$$
\begin{aligned}
Y_1 &= X_1 && (70)\\
Y_2 &= X_2 && (71)\\
&\quad (X_V = X_1 + X_2) && (72)\\
Y_3 &= X_V + X_{3|V} = X_1 + X_2 + X_{3|X_1,X_2} && (73)\\
Y_4 &= X_V + X_{4|V} = X_1 + X_2 + X_{4|X_1,X_2} && (74)\\
Y_5 &= X_V + X_{5|V} = X_1 + X_2 + X_{5|X_1,X_2} && (75)
\end{aligned}
$$

The calculation of the variances is trivial. As far as the covariances we have

$$
\begin{aligned}
\mathrm{Cov}[Y_1, Y_2] &= 0 && (76)\\
\mathrm{Cov}[Y_1, Y_i] &= \sigma_1^2 \qquad (i > 2) && (77)\\
\mathrm{Cov}[Y_2, Y_i] &= \sigma_2^2 \qquad (i > 2) && (78)\\
\mathrm{Cov}[Y_i, Y_j] &= \sigma_1^2 + \sigma_2^2 \quad (i > 2,\ j > 2) && (79)
\end{aligned}
$$

This is then the covariance matrix of interest, limited to the five variables shown in the figure (and than it is easy to continue):

$$
\boldsymbol{V} = \begin{pmatrix}
\sigma_1^2 & 0 & \sigma_1^2 & \sigma_1^2 & \sigma_1^2 \\
0 & \sigma_2^2 & \sigma_2^2 & \sigma_2^2 & \sigma_2^2 \\
\sigma_1^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + \sigma_{3|1,2}^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 + \sigma_2^2 \\
\sigma_1^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + \sigma_{4|1,2}^2 & \sigma_1^2 + \sigma_2^2 \\
\sigma_1^2 & \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 + \sigma_2^2 & \sigma_1^2 + \sigma_2^2 + \sigma_{5|1,2}^2
\end{pmatrix} \qquad (80)
$$

From such an interesting matrix we can expect interesting results, useful to *train our intuition*. But before analyzing some cases, as done in the previous section, let us make the exercise to build up the covariance matrix in a different way. The transformation rules (70)-(75) can be rewritten using the transformation matrix

$$
\boldsymbol{C} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 0 & 1
\end{pmatrix} \qquad (81)
$$

to be applied to the diagonal matrix of the independent variables,

$$
\boldsymbol{V}_0 = \begin{pmatrix}
\sigma_1^2 & 0 & 0 & 0 & \\
0 & \sigma_2^2 & 0 & 0 & 0 \\
0 & 0 & \sigma_{3|1,2}^2 & 0 & \\
0 & 0 & 0 & \sigma_{4|1,2}^2 & 0 \\
0 & 0 & 0 & 0 & \sigma_{5|1,2}^2
\end{pmatrix} \qquad (82)
$$

Applying then the transformation rule of the covariance matrix we reobtain the above result – an implementation in R will be shown in the next subsection.

6.1 Numerical examples

Let set up the covariance matrix for 5 possible 'observations'

```
> n=7; muX1=0; sigmaX1=10; muZ=0; sigmaZ=1      # set parameters
> mu <- c(muX1, muZ, rep(muX1+muZ, n-2))        # set expected values
> ( sigma <- c(sigmaX1, sigmaZ, rep(1,n-2)) )   # standard deviations
[1] 10  1  1  1  1  1  1
> V <- matrix(rep(0, n*n), c(n,n))              # cov matr  # step 0
> V[(1:n)[-2], (1:n)[-2]] <-  sigma[1]^2                  # step 1
> V[(2:n), (2:n)]         <-  V[(2:n), (2:n)] + sigma[2]^2  # step 2
> diag(V)[3:n] <- diag(V)[3:n] + sigma[3:n]^2             # step 3
> V
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  100    0  100  100  100  100  100
[2,]    0    1    1    1    1    1    1
[3,]  100    1  102  101  101  101  101
[4,]  100    1  101  102  101  101  101
[5,]  100    1  101  101  102  101  101
[6,]  100    1  101  101  101  102  101
[7,]  100    1  101  101  101  101  102
> (su <- sqrt(diag(V)))
[1] 10.0000  1.0000 10.0995 10.0995 10.0995 10.0995 10.0995
> round( V/outer(su,su), 4)
       [,1]  [,2]   [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 1.0000 0.000 0.9901 0.9901 0.9901 0.9901 0.9901
[2,] 0.0000 1.000 0.0990 0.0990 0.0990 0.0990 0.0990
[3,] 0.9901 0.099 1.0000 0.9902 0.9902 0.9902 0.9902
[4,] 0.9901 0.099 0.9902 1.0000 0.9902 0.9902 0.9902
[5,] 0.9901 0.099 0.9902 0.9902 1.0000 0.9902 0.9902
[6,] 0.9901 0.099 0.9902 0.9902 0.9902 1.0000 0.9902
[7,] 0.9901 0.099 0.9902 0.9902 0.9902 0.9902 1.0000
```

Let us also show the **alternative way to build up the covariance matrix**

```
> C <- matrix(rep(0, n*n), c(n,n))              # transf. matrix
> C[,1] <- c(1, 0, rep(1, n-2))
> C[,2] <- c(0, rep(1, n-1))
> diag(C) <- rep(1, n)
> C
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    1    0    0    0    0    0    0
[2,]    0    1    0    0    0    0    0
[3,]    1    1    1    0    0    0    0
[4,]    1    1    0    1    0    0    0
[5,]    1    1    0    0    1    0    0
[6,]    1    1    0    0    0    1    0
[7,]    1    1    0    0    0    0    1
> V0 <- matrix(rep(0, n*n), c(n,n))     # initial diagonal matrix
> diag(V0) <-  sigma^2
```

```
> ( V <- C %*% V0 %*% t(C) )            # joint covariance matrix
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  100    0  100  100  100  100  100
[2,]    0    1    1    1    1    1    1
[3,]  100    1  102  101  101  101  101
[4,]  100    1  101  102  101  101  101
[5,]  100    1  101  101  102  101  101
[6,]  100    1  101  101  101  102  101
[7,]  100    1  101  101  101  101  102
```

As we see the result is identical to that obtained setting the elements 'by hand'.

Then let us now repeat the steps previously followed without systematic offset.

6.1.1 Condition on $X_1 = 2$ ("known true value")

```
> ( mu.c <- c(2, rep(NA, n-1)) )
[1]  2 NA NA NA NA NA NA
> ( out <- norm.mult.cond(mu, V, mu.c) )
$mu
[1] 2 0 2 2 2 2 2

$V
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    0    0    0    0    0    0    0
[2,]    0    1    1    1    1    1    1
[3,]    0    1    2    1    1    1    1
[4,]    0    1    1    2    1    1    1
[5,]    0    1    1    1    2    1    1
[6,]    0    1    1    1    1    2    1
[7,]    0    1    1    1    1    1    2
> ( out.s <- sqrt(diag(out$V)) )
[1] 0.000000 1.000000 1.414214 1.414214 1.414214 1.414214 1.414214
> round( out$V / outer(out.s, out.s), 3)
     [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,]  NaN   NaN   NaN   NaN   NaN   NaN   NaN
[2,]  NaN 1.000 0.707 0.707 0.707 0.707 0.707
[3,]  NaN 0.707 1.000 0.500 0.500 0.500 0.500
[4,]  NaN 0.707 0.500 1.000 0.500 0.500 0.500
[5,]  NaN 0.707 0.500 0.500 1.000 0.500 0.500
[6,]  NaN 0.707 0.500 0.500 0.500 1.000 0.500
[7,]  NaN 0.707 0.500 0.500 0.500 0.500 1.000
```

The condition on the 'true value' changes the values of the observables to its value, but it does not affect the offset, which has a role in the uncertainty of the future observations as well in their correlation. In fact, contrary to the case see in the previous section without uncertain offset, they are not any longer independent. They would become independent if also the offset were known (try for example with "`mu.c <- c(2, 0, rep(NA, n-2))`" to see the difference, or even better with "`mu.c <- c(2, 1, rep(NA, n-2))`").

6.1.2 Condition on $X_3 = 2$ ("single observation")

```
> ( mu.c <- c(NA, NA, 2, rep(NA, n-3)) )
[1] NA NA  2 NA NA NA NA
> out <- norm.mult.cond(mu, V, mu.c)
> round( out$mu, 4)
[1] 1.9608 0.0196 2.0000 1.9804 1.9804 1.9804 1.9804
> round( out$V, 4)
        [,1]    [,2] [,3]   [,4]   [,5]   [,6]   [,7]
[1,]  1.9608 -0.9804    0 0.9804 0.9804 0.9804 0.9804
[2,] -0.9804  0.9902    0 0.0098 0.0098 0.0098 0.0098
[3,]  0.0000  0.0000    0 0.0000 0.0000 0.0000 0.0000
[4,]  0.9804  0.0098    0 1.9902 0.9902 0.9902 0.9902
[5,]  0.9804  0.0098    0 0.9902 1.9902 0.9902 0.9902
[6,]  0.9804  0.0098    0 0.9902 0.9902 1.9902 0.9902
[7,]  0.9804  0.0098    0 0.9902 0.9902 0.9902 1.9902
> round( out.s <- sqrt(diag(out$V)), 4 )
[1] 1.4003 0.9951 0.0000 1.4107 1.4107 1.4107 1.4107
> round( out$V / outer(out.s, out.s), 3)
       [,1]   [,2] [,3]  [,4]  [,5]  [,6]  [,7]
[1,]  1.000 -0.704  NaN 0.496 0.496 0.496 0.496
[2,] -0.704  1.000  NaN 0.007 0.007 0.007 0.007
[3,]    NaN    NaN  NaN   NaN   NaN   NaN   NaN
[4,]  0.496  0.007  NaN 1.000 0.498 0.498 0.498
[5,]  0.496  0.007  NaN 0.498 1.000 0.498 0.498
[6,]  0.496  0.007  NaN 0.498 0.498 1.000 0.498
[7,]  0.496  0.007  NaN 0.498 0.498 0.498 1.000
```

To understand the result we need to compare it with the case without uncertainty uncertainty. In that case we had $X_1 = 1.98$. Now we have $X_1 = 1.96$. The difference, although practically irrelevant, is conceptually important. It corresponds in fact to the expected value of the offset (precisely 0.0196). Indeed, the role of the observation is to give us an information about $X_1 + X_2$, sum of the true value and the offset. The fact that we use the observations to update our knowledge on the true value is simply because the offset is a priori better known that the true value, as it is well understood by experienced physicists: if the calibration is poor the instrument cannot be used for 'measurements'. Note also the correlation that now appears between X_1 and X_2, and in particular its negative sign: the value of the true value could increase at the 'expenses' of the offset, and the other way around.

6.1.3 Condition on $X_3 = 2$ and $X_4 = 1$ ("two observations")

```
> ( mu.c <- c(NA, NA, 2, 1, rep(NA, n-4)) )
[1] NA NA  2  1 NA NA NA
> out <- norm.mult.cond(mu, V, mu.c)
> round( out$mu, 4)
[1] 1.4778 0.0148 2.0000 1.0000 1.4926 1.4926 1.4926
```

```
> round( out$V, 4)
        [,1]    [,2] [,3] [,4]   [,5]   [,6]   [,7]
[1,]  1.4778 -0.9852    0    0 0.4926 0.4926 0.4926
[2,] -0.9852  0.9901    0    0 0.0049 0.0049 0.0049
[3,]  0.0000  0.0000    0    0 0.0000 0.0000 0.0000
[4,]  0.0000  0.0000    0    0 0.0000 0.0000 0.0000
[5,]  0.4926  0.0049    0    0 1.4975 0.4975 0.4975
[6,]  0.4926  0.0049    0    0 0.4975 1.4975 0.4975
[7,]  0.4926  0.0049    0    0 0.4975 0.4975 1.4975
> round( out.s <- sqrt(diag(out$V)), 4 )
[1] 1.2157 0.9951 0.0000 0.0000 1.2237 1.2237 1.2237
> round( out$V / outer(out.s, out.s), 3)
       [,1]   [,2] [,3] [,4]  [,5]  [,6]  [,7]
[1,]  1.000 -0.814  NaN  NaN 0.331 0.331 0.331
[2,] -0.814  1.000  NaN  NaN 0.004 0.004 0.004
[3,]    NaN    NaN  NaN  NaN   NaN   NaN   NaN
[4,]    NaN    NaN  NaN  NaN   NaN   NaN   NaN
[5,]  0.331  0.004  NaN  NaN 1.000 0.332 0.332
[6,]  0.331  0.004  NaN  NaN 0.332 1.000 0.332
[7,]  0.331  0.004  NaN  NaN 0.332 0.332 1.000
```

The only new effect we observe is the increase (in module) of the correlation coefficient between true value and offset. This is due to the fact that the increased number of observation has increased the constrain between the two quantities. It will increase more if we use furher observations, for example conditioning on "`mu.c <- c(NA, NA, 2, 1, 1.5, 2.2, 0.5)`", or decreasing the standard deviations $\sigma_{i|1,2}$. For example if we set all $\sigma_{i|1,2}$ to 0.1, the same conditioning on X_3 and X_3 would produce a correlation coefficient of -0.9975.

6.1.4 "Ricalibration of the offset" ($X_1 = 2$; $X_3 = 2$, $X_4 = 1$)

What happens if we instead fix the value of the true value and some values of the observables? In this case we update our information on the offset. Let us see the case in which we fix the value of the true value at 2, and the average of the two observations at 1.5.

```
> ( mu.c <- c(2, NA, 2, 1, rep(NA, n-4)) )
[1]  2 NA  2  1 NA NA NA
> out <- norm.mult.cond(mu, V, mu.c)
> round( out$mu, 4)
[1]  2.0000 -0.3333  2.0000  1.0000  1.6667  1.6667  1.6667
> round( out$V, 4)
     [,1]   [,2] [,3] [,4]   [,5]   [,6]   [,7]
[1,]    0 0.0000    0    0 0.0000 0.0000 0.0000
[2,]    0 0.3333    0    0 0.3333 0.3333 0.3333
[3,]    0 0.0000    0    0 0.0000 0.0000 0.0000
[4,]    0 0.0000    0    0 0.0000 0.0000 0.0000
[5,]    0 0.3333    0    0 1.3333 0.3333 0.3333
[6,]    0 0.3333    0    0 0.3333 1.3333 0.3333
```

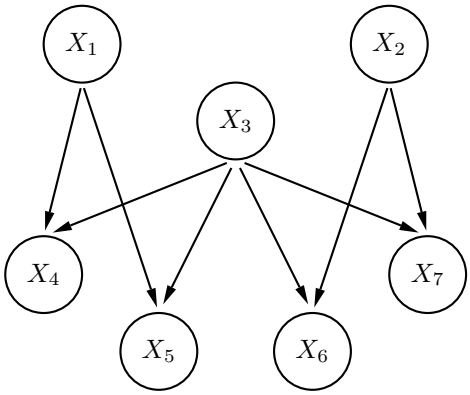


Figure 9: Basic models of joint probabilities

```
[7,]    0 0.3333    0    0 0.3333 0.3333 1.3333
> round( out.s <- sqrt(diag(out$V)), 4 )
[1] 0.0000 0.5774 0.0000 0.0000 1.1547 1.1547 1.1547
> round( out$V / outer(out.s, out.s), 3)
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  NaN  NaN  NaN  NaN  NaN  NaN  NaN
[2,]  NaN  1.0  NaN  NaN 0.50 0.50 0.50
[3,]  NaN  NaN  NaN  NaN  NaN  NaN  NaN
[4,]  NaN  NaN  NaN  NaN  NaN  NaN  NaN
[5,]  NaN  0.5  NaN  NaN 1.00 0.25 0.25
[6,]  NaN  0.5  NaN  NaN 0.25 1.00 0.25
[7,]  NaN  0.5  NaN  NaN 0.25 0.25 1.00
```

As a result, the expected value of the offset becomes -0.33, with a standard deviation of 0.58, against the (possible) intuitive guess of -0.5 (i.e $1.5-2.0$) with a standard uncertainty of 0.71 (i.e. $1/\sqrt{2}$). The reason is that our prior knowledge on the offset had a standard uncertainty of 1, that has to be taken into account. Indeed it can be easily checked that the 'intuitive' result would have been recovered if we had a very large uncertainty ($\sigma_2 \to \infty$). In fact -0.33 is the weighted average of the initial value 0 and -0.5, with weights equal to 1 and 2. The reason is that result based on reconditioning provides automatically the rule of the weighted average with 'inverse of the variances', where the 'variance' associated to -0.5 would be that obtained if the prior knowledge on the offset was irrelevant (i.e. $\sigma_2 \to \infty$).

7 Measuring two quantity with the same instrument affected by an offset uncertainty

Another interesting issue, very common in experimental physics, is when we make several measurements on homogeneous quantities using the same instrument that, as all instru-

ments, has unavoidable uncertainty in the calibration. The situation is sketched in the diagram of Fig. 9, where X_1 and X_2 are the true values, X_3 the common offset, X_4 and X_5 the readings when the instrument is applied to X_1, X_6 and X_7 the readings when the instrument is applied to X_2.

From this model we can easily build the transformation matrix $\boldsymbol{C}$

$$\boldsymbol{C} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{83}$$

(for example its says that row 6 depends on X_2, X_3 and $X_{6|2,3}$). Applying it to the starting diagonal matrix (X_1, X_2 and X_3 are *initially independent*, and also $X_{i|2,3}$ and so on are conditionally independent) we get the covariance matrix of the joint multivariate normal of interest:

$$\boldsymbol{V} = \begin{pmatrix} \sigma_1^2 & 0 & 0 & \sigma_1^2 & \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 & 0 & 0 & \sigma_2^2 & \sigma_2^2 \\ 0 & 0 & \sigma_3^2 & \sigma_3^2 & \sigma_3^2 & \sigma_3^2 & \sigma_3^2 \\ \sigma_1^2 & 0 & \sigma_3^2 & \sigma_1^2+\sigma_3^2+\sigma_{4|1,3}^2 & \sigma_1^2+\sigma_3^2 & \sigma_3^2 & \sigma_3^2 \\ \sigma_1^2 & 0 & \sigma_3^2 & \sigma_1^2+\sigma_3^2 & \sigma_1^2+\sigma_3^2+\sigma_{5|1,3}^2 & \sigma_3^2 & \sigma_3^2 \\ 0 & \sigma_2^2 & \sigma_3^2 & \sigma_3^2 & \sigma_3^2 & \sigma_2^2+\sigma_3^2+\sigma_{6|2,3}^2 & \sigma_2^2+\sigma_3^2 \\ 0 & \sigma_2^2 & \sigma_3^2 & \sigma_3^2 & \sigma_3^2 & \sigma_2^2+\sigma_3^2 & \sigma_2^2+\sigma_3^2+\sigma_{7|2,3}^2 \end{pmatrix}$$

This is a very interesting covariance matrix and we leave the reader the pleasure of exploiting all possibilities. Here we just show a numerical example, with parameters similar to the ones used before for a better understanding, and just discuss a single case of conditioning.

```
> n=7; muX1=0; sigmaX1=10; muX2=0; sigmaX2=10;  # set parameters
> muZ=0; sigmaZ=1
> mu <- c(muX1, muX2, muZ, rep(muX1+muZ,2), rep(muX2+muZ,2)) # set expected values
> ( sigma <- c(sigmaX1, sigmaX1, sigmaZ, rep(1, n-3)) )   # standard deviations
[1] 10 10  1  1  1  1  1
> C <- matrix(rep(0, n*n), c(n,n))      # tranformation matrix
> diag(C) <- rep(1, n)
> C[4,] <- c(1, 0, 1, 1, 0, 0, 0)
> C[5,] <- c(1, 0, 1, 0, 1, 0, 0)
> C[6,] <- c(0, 1, 1, 0, 0, 1, 0)
> C[7,] <- c(0, 1, 1, 0, 0, 0, 1)
> C
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]    1    0    0    0    0    0    0
```

```
[2,]    0    1    0    0    0    0    0
[3,]    0    0    1    0    0    0    0
[4,]    1    0    1    1    0    0    0
[5,]    1    0    1    0    1    0    0
[6,]    0    1    1    0    0    1    0
[7,]    0    1    1    0    0    0    1
> V0 <- matrix(rep(0, n*n), c(n,n))     # covariance matrix
> diag(V0) <-  sigma^2
> V <- C %*% V0 %*% t(C)
> V
     [,1] [,2] [,3] [,4] [,5] [,6] [,7]
[1,]  100    0    0  100  100    0    0
[2,]    0  100    0    0    0  100  100
[3,]    0    0    1    1    1    1    1
[4,]  100    0    1  102  101    1    1
[5,]  100    0    1  101  102    1    1
[6,]    0  100    1    1    1  102  101
[7,]    0  100    1    1    1  101  102
> su <- sqrt(diag(V))                   # standard uncertainties
> round( V/outer(su,su), 4)             # correlation matrix
       [,1]   [,2]  [,3]   [,4]   [,5]   [,6]   [,7]
[1,] 1.0000 0.0000 0.000 0.9901 0.9901 0.0000 0.0000
[2,] 0.0000 1.0000 0.000 0.0000 0.0000 0.9901 0.9901
[3,] 0.0000 0.0000 1.000 0.0990 0.0990 0.0990 0.0990
[4,] 0.9901 0.0000 0.099 1.0000 0.9902 0.0098 0.0098
[5,] 0.9901 0.0000 0.099 0.9902 1.0000 0.0098 0.0098
[6,] 0.0000 0.9901 0.099 0.0098 0.0098 1.0000 0.9902
[7,] 0.0000 0.9901 0.099 0.0098 0.0098 0.9902 1.0000
```

Now let us assume we have applied our instrument once on X_1 and once on X_2, obtaining the readings $X_4 = 1$ and $X_6 = 2$, respectively. Here is how our knowledge is updated:

```
> ( mu.c <- c(rep(NA, 3), 1, NA, 2, NA) )       # conditioning
[1] NA NA NA  1 NA  2 NA
> out <- norm.mult.cond(mu, V, mu.c)
> round( out$mu, 4)
[1] 0.9613 1.9514 0.0291 1.0000 0.9904 2.0000 1.9805
> round( out$V, 4)
        [,1]    [,2]    [,3] [,4]    [,5] [,6]    [,7]
[1,]  1.9514  0.9613 -0.9709    0  0.9805    0 -0.0096
[2,]  0.9613  1.9514 -0.9709    0 -0.0096    0  0.9805
[3,] -0.9709 -0.9709  0.9806    0  0.0097    0  0.0097
[4,]  0.0000  0.0000  0.0000    0  0.0000    0  0.0000
[5,]  0.9805 -0.0096  0.0097    0  1.9902    0  0.0001
[6,]  0.0000  0.0000  0.0000    0  0.0000    0  0.0000
[7,] -0.0096  0.9805  0.0097    0  0.0001    0  1.9902
> round( out.s <- sqrt(diag(out$V)), 4 )
[1] 1.3969 1.3969 0.9902 0.0000 1.4107 0.0000 1.4107
> round( out$V / outer(out.s, out.s), 3)
```

```
       [,1]   [,2]   [,3] [,4]   [,5] [,6]   [,7]
[1,]  1.000  0.493 -0.702  NaN  0.498  NaN -0.005
[2,]  0.493  1.000 -0.702  NaN -0.005  NaN  0.498
[3,] -0.702 -0.702  1.000  NaN  0.007  NaN  0.007
[4,]    NaN    NaN    NaN  NaN    NaN  NaN    NaN
[5,]  0.498 -0.005  0.007  NaN  1.000  NaN  0.000
[6,]    NaN    NaN    NaN  NaN    NaN  NaN    NaN
[7,] -0.005  0.498  0.007  NaN  0.000  NaN  1.000
```

As expected, $X_4 = 1$ sets essentially to 1 the true value X_1 and the 'future' – or not yet known! – reading X_5. Similarly, X_6 sets essentially to 2 X_2 and X_7. (The difference from the exact value of 1 and 2, respectively, is due – let us repeat it once again – to the fact that we use, for didascalic purposes, initial standard uncertainties σ_1 and σ_2 'relatively small', while the uncertainty on the common offset is 'relatively large'.) The most interest part of the result is the 3×3 upper left part of the resulting correlation matrix, which we repeat here. As we have learned in the previous section, the value of the offset gets anticorrelated to the true values. Moreover the two true values get **positively correlated**, as expected: a part of our uncertainty on them is due the imprecise knowledge of the offset, which then affect both values in the same direction.

8 Memento

visualizzazione variabili multivariate con monte Carlo

effetto su X1 di condizionamenti separati su osservabili ???

9 Propagation of evidence – some general remarks

Let us take again the diagrams ('*graphs*') which describe two observations from the same true value and one observation resulting from a true value and a systematic effect. They are show again in Fig. 10, labelled with names related to the direction of the 'causation' arrows, which *diverge* from a single *node* or *converge* towards a single node. The physical interpretation is that, as we have already seen, of a single *cause* producing two *effects*, or two *causes* responsible of a single effect below each graph we have also added the covariance matrix which characterize them. Fore completeness we have added in the figure also graph in which the effect X_2 is itself cause of another effect (*serial connection*).

Sticking to the simple linear models we are dealing with, the trasformation rules of the graph characterize by a serial connection are the following:

$$Y_1 = X_1 \tag{84}$$

$$Y_2 = X_2|_{X_1} + X_1 \tag{85}$$

$$Y_3 = X_3|_{X_2} + Y_2 = X_3|_{X_2} + X_2|_{X_1} + X_1\,, \tag{86}$$

from which the joint covariance matrix reported below the diagram follows.

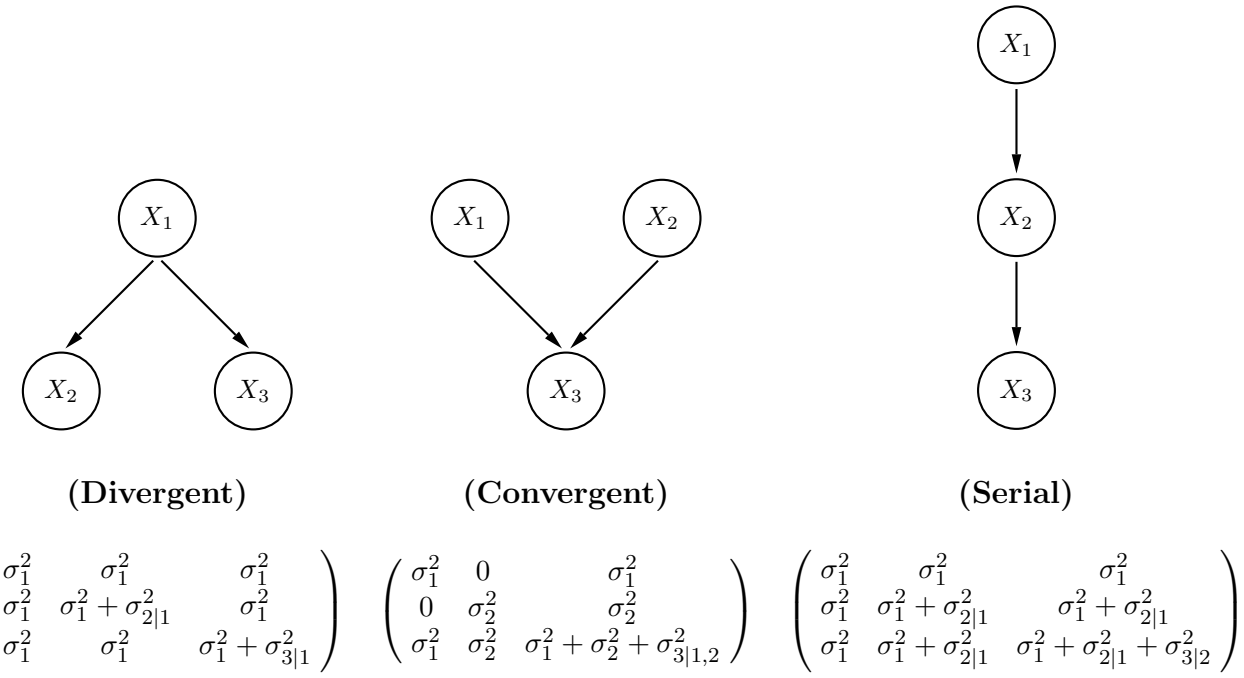


Figure 10: Basic models of joint probabilities

Analyzing the covariance matrix of the graphs with divergent and serial connections we see that the variables are fully correlated: any hypothesis on any of the three variables changes the pdf of the other two. This is clear if we start from X_1, as indicated by the arrows, but it is also true if we start from X_2 or X_3 as indicated by the dashed arrows in figure 11.

Instead, in the convergent graph X_1 and X_2 are independent: why should the physical quantity we are goind to measure shoud depend on a calibration constant of our detector? And the other way around.[15] But we have already seen in the examples that if we observe X_3, then X_1 and X_2 become anticorrelated.

The effect of a condition ('*instantiation*') of one variable to the rest of the *network* is very interesting, also for its practical applications, because it allows to decompose a large network in subnetworks.

[15]In reality it is not impossible to think to sophisticated cases in which the value of the physical constant we going to measure could influence our beliefs about the performance of the detector. Let us say that the assumptions of independence between 'true value' and instrument offset is more than reasonable in most, if not all, experimental cases.

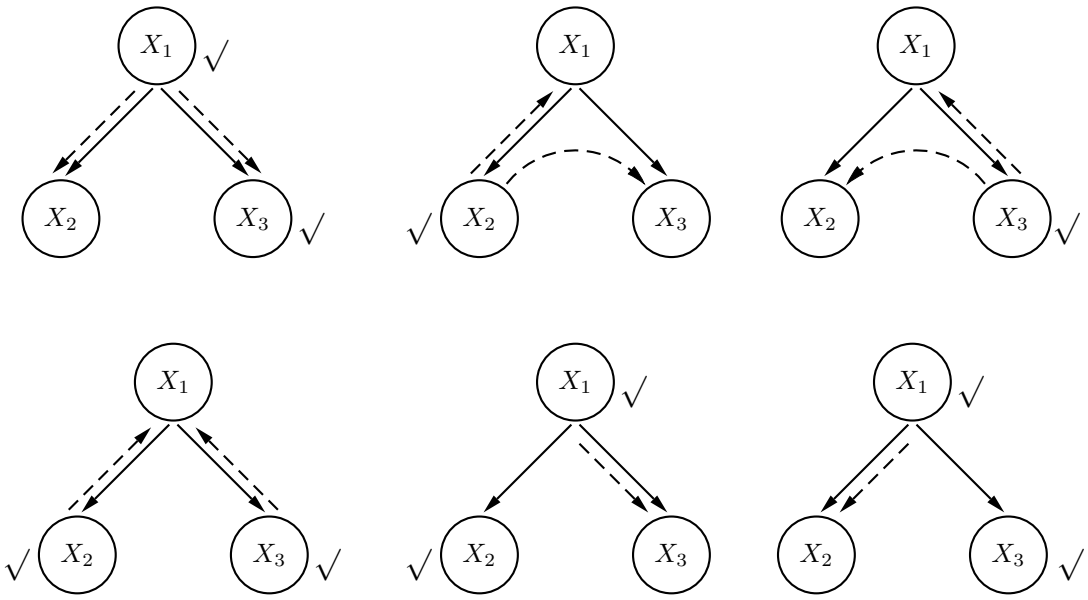


Figure 11: Divergent connection with some with 'evidence' got in some of the variables ('instatiated nodes'). The dashed arrows show the 'flow of evidence', i.e. how uncertainty is updated troughtough the 'network'.

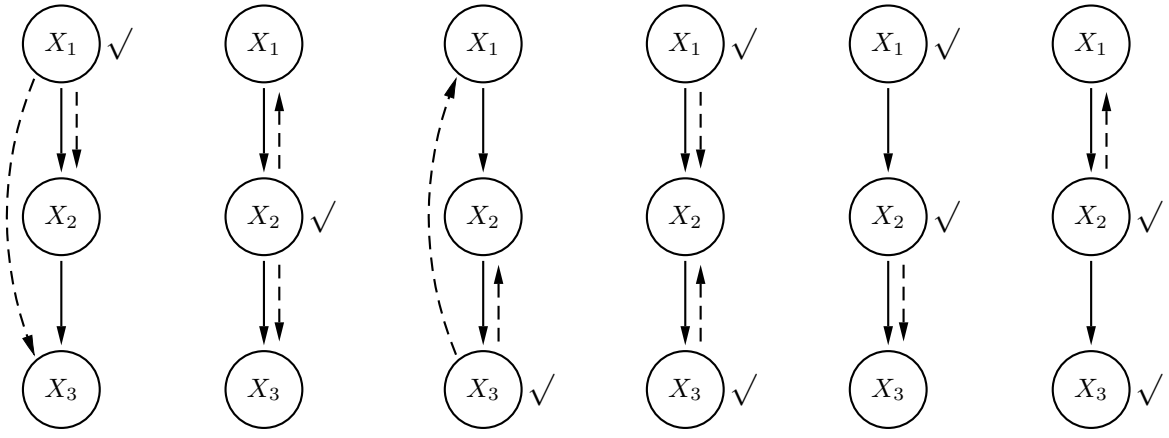


Figure 12: As Fig. 11 for a serial connection.

9.1 Diverging connection

We have already seen in the numerical examples of subsection 5.1 that if we condition on a value of X_1, then X_2 and X_3 become independent, and the physical reason was very easy to understand. This is a general property of divergent graphs, usually stated referring to *parents* and *children*: in a divergent graph, if a parent is instantiated, the children become independent, i.e. *evidence does not flow from one child to the other* ('an instantiated parent blocks evidence flow among children' – we assume that there is no other connection among them!).

Let us make the exercise to calculate the covariance matrix of X_2 and X_3 given X_1. To use the compact formula shown above, we need to rewrite the three variable in a compact form, thus defining $\boldsymbol{Y}_1 = \{X_1\}$ and $\boldsymbol{Y}_2 = \{X_2, X_3\}$. Then it is convenient to rewrite formula (39) swapping the indices, though obtaining:

$$\boldsymbol{V}\left[\boldsymbol{Y}_2|_{\boldsymbol{Y}_1}\right] = \boldsymbol{V}_{22} - \boldsymbol{V}_{21}\,\boldsymbol{V}_{11}^{-1}\,\boldsymbol{V}_{12}\,, \tag{87}$$

in which we recognize

$$\boldsymbol{V}_{22} = \begin{pmatrix} \sigma_1^2 + \sigma_{2|1}^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 + \sigma_{3|1}^2 \end{pmatrix} \tag{88}$$

$$\boldsymbol{V}_{21} = \begin{pmatrix} \sigma_1^2 \\ \sigma_1^2 \end{pmatrix} \tag{89}$$

$$\boldsymbol{V}_{11} = \sigma_1^2 \tag{90}$$

$$\boldsymbol{V}_{11}^{-1} = \frac{1}{\sigma_1^2} \tag{91}$$

$$\boldsymbol{V}_{21} = \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \end{pmatrix} \tag{92}$$

It follows

$$\boldsymbol{V}_{21}\,\boldsymbol{V}_{11}^{-1}\,\boldsymbol{V}_{12} = \begin{pmatrix} \sigma_1^2 \\ \sigma_1^2 \end{pmatrix} \cdot \frac{1}{\sigma_1^2} \cdot \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_1^2 \\ \sigma_1^2 & \sigma_1^2 \end{pmatrix} \tag{93}$$

and hence

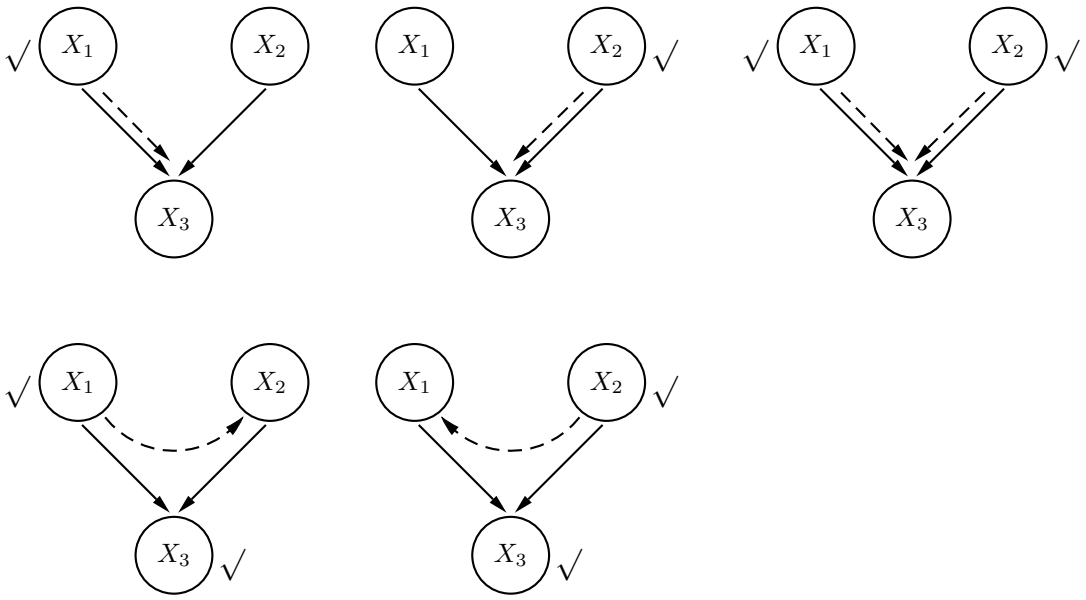
$$\boldsymbol{V}\left[\boldsymbol{Y}_2|_{\boldsymbol{Y}_1}\right] = \begin{pmatrix} \sigma_{2|1}^2 & 0 \\ 0 & \sigma_1^2 + \sigma_{3|1}^2 \end{pmatrix} \tag{94}$$

As expected, the exercise shows that X_2 and X_3 are independent.

9.2 Converging connection

9.3 Serial connection

Pearl



As Fig. 11 for a converging connection.

10 Conclusions

11 Appendix: probabilities inside contour ellipses and contour ellipsoids

References

[1] R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`.

[2] C.F. Gauss, *"Theoria motus corporum coelestium in sectionibus conicis solem ambientum"*, Hamburg 1809, n.i 172–179; reprinted in Werke, Vol. 7 (Gota, Göttingen, 1871), pp 225–234.
(See e.g. in `http://www.roma1.infn.it/~dagos/history/GaussMotus/index.html`)

[3] M. Eaton, *Multivariate Statistics : A Vector Space Approach*, John Wiley and Sons 1983 (available at `http://projecteuclid.org/`).

[4] G. D'Agostini, *Bayesian reasoning in data analysis – a critical introduction*, World Scientific 2003.