Fits, and especially linear fits, with errors on both axes, extra variance of the data points and other complications

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

Abstract

The aim of this paper, triggered by some discussions in the astrophysics community raised by astro-ph/0508529, is to introduce the issue of 'fits' from a probabilistic perspective (also known as Bayesian), with special attention to the construction of model that describes the 'network of dependences' (a Bayesian network) that connects experimental observations to model parameters and upon which the probabilistic inference relies. The particular case of linear fit with errors on both axes and extra variance of the data points around the straight line (i.e. not accounted by the experimental errors) is shown in detail. Some questions related to the use of linear fit formulas to log-linearized exponential and power laws are also sketched, as well as the issue of systematic errors.

Preamble

This paper, based on things already written somewhere with the addition of some details from lectures, contains nothing or little especially new. Even the main 'result', summarized in Eq.(35) and that I hope will contribute to set down the questions raised by astro-ph/0508529 [1], is just a simple extension of Eq. (8.33) of Ref. [2]. Therefore the debated question could be dismissed with a paper even shorter than astro-ph/0508529. Nevertheless, I have taken the opportunity to reorganize old material for the benefit of my students, and I post these pages hoping they could be of some utility to those who wish to understand what there is behind formulas.

1 Introduction

A common task in data analysis is to 'determine', on the basis of experimental observations, the values of the parameters of a model that relates physical quantities. This procedure is usually associated to names like 'fit' and 'regression', and to *principles*, like 'least squares' or 'maximum likelihood' (with variants). I prefer, as many others belonging to a still small minority, to approach the problem from more fundamental probabilistic 'first principles', that are indeed the fundamental rules of probability theory. This approach is also called 'Bayesian' because of the central role played by Bayes' theorem in the process of learning from data, as we shall see in a while (for a critical introduction to the Bayesian approach see Ref. [2] and references therein). In practice this means that we rank in probability hypotheses and numerical values about which we are not certain. This is rather intuitive and it is indeed the natural way physicists reason (see e.g. Ref. [3] and references therein), though we have been taught a peculiar view of probability that does not allow us to make the reasonings we intuitively do and that we are going to use here.

In the so called Bayesian approach the issue of 'fits' takes the name of *parametric inference*, in the sense we are interested in inferring the parameters of a model that relates 'true' values. The outcome of the inference is an uncertain knowledge of parameters, whose possible values are ranked using the language and the tools of probability theory. As it *can only be* (see e.g. Ref. [2] for extensive discussions), the resulting inference depends on the inferential model and on previous knowledge about the possible values the model parameters can take (though this last dependence is usually rather weak if the inference is based on a 'large' number of observations). It is then important to state clearly the several assumptions that enter the data analysis. I hope this paper does it with the due care – and I apologize in advance for some pedantry and repetitions. The main message I would like to convey is that nowadays it is much more important to build up the model that describes at best the physics case than to obtain simple formulae for the 'best estimates' and their uncertainty. This is because, thanks to the extraordinary progresses of applied mathematics and computing power, in most cases the calculation of the integrals that come from a straight application of the probability theory does not require any longer titanic efforts. Building up the correct model is then equivalent, in most cases, to have solved the problem.

The paper is organized as follows. In Section 2 the inferential approach is introduced from scratch, only assuming the multivariate extensions of the following well known formulas[1]

$$f(x, y \,|\, I) = f(x \,|\, y, I) \cdot f(y \,|\, I) \tag{1}$$

$$f(x \,|\, I) = \int f(x, y \,|\, I)\, dy\,. \tag{2}$$

We show how to build the general model, and how this evolves as soon as the several hypotheses of the model are introduced (independence, normal error functions, linear dependence between

[1]The meaning of the overall conditioning I will be clarified later. Note that, in order to simplify the notation, the generic symbol $f(\,)$ is used to indicate all probability density functions, though they might refer to different variables and have different mathematical expressions. In particular, the order of the arguments is irrelevant, in the sense that $f(x, y \,|\, I)$ stands for 'joint probability density function of x and y under condition I', and therefore it could be also indicated by $f(y, x \,|\, I)$. For the same reason, the indexes of sums and products and the extremes of the integrals are usually omitted, implying they extend to all possible values of the variables.

true values, vague priors). The graphical representation of the model in terms of the so called 'Bayesian networks' is also shown, the utility of which will become self-evident. The case of linear fit with errors on both axes is then summarized in Section 3, and the approximate solution for the non-linear case is sketched in Section 4. The extra variability of the data is modeled in Section 5, first in general and then in the simple case of the linear fit. The interpretation of the inferential result is discussed in Section 6, in which approximated methods to calculate the fit summaries (expected values and variance of the parameters) are shown. Finally, some comments on the not-trivial issues related to the use of linear fit formulas to infer the parameters of exponential and power laws are given in Section 7. Section 8 shows how to extend the model to include systematic errors, and some simple formulas to take into account offset and scale systematic errors in the case of linear fits will be provided. The paper ends with some conclusions and some comments about the debate that has triggered it.

2 Probabilistic parametric inference from a set of data points with errors on both axes

Let us consider a 'law' that relates the 'true' values of two quantities, indicated here by μ_x and μ_y:

$$\mu_y = \mu_y(\mu_x; \boldsymbol{\theta}) \,, \tag{3}$$

where $\boldsymbol{\theta}$ stands for the parameters of the law, whose number is M. In the linear case Eq. (3) reduces to

$$\mu_y \;\; = \;\; m\, \mu_x + c \tag{4}$$

i.e. $\boldsymbol{\theta} = \{m, c\}$ and $M = 2$. As it is well understood, because of 'errors' we do not observe directly μ_x and μ_y, but experimental quantities[2] x and y that might differ, on an event by event basis, from μ_x and μ_y. The outcome of the 'observation' (see footnote 2) x_i for a given μ_{x_i} (analogous reasonings apply to y_i and μ_{y_i}) is modeled by an error function $f(x_i \,|\, \mu_{x_i}, I)$, that is indeed a probability density function (pdf) conditioned by μ_{x_i} and the 'general state of knowledge' I. The latter stands for all background knowledge behind the analysis, that is what for example makes us to believe the relation $\mu_y = \mu_y(\mu_x; \boldsymbol{\theta})$, the particular mathematical expressions for $f(x_i \,|\, \mu_{x_i}, I)$ and $f(y_i \,|\, \mu_{y_i}, I)$, and so on. Note that the shape of the error function might depend on the value of μ_{x_i}, as it happens if the detector does not respond the same way to different solicitations. A usual *assumption* is that *errors are normally distributed*, i.e.

$$x_i \;\; \sim \;\; \mathcal{N}(\mu_{x_i}, \sigma_{x_i}) \tag{5}$$

$$y_i \;\; \sim \;\; \mathcal{N}(\mu_{y_i}, \sigma_{y_i}) \,, \tag{6}$$

where the symbol '$\sim$' stands for 'is described by the distribution' (or 'follows the distribution'), and where we still leave the possibility that the standard deviations, that we consider known,

[2]These quantities might also be summaries of the data. I.e. they are either directly observed numbers, like readings on scales, or quantities calculated from direct observations, like averages or other 'statistics' based on partial analysis of the data. It is implicit that when summaries are used, instead of direct observations, the analyzer is somewhat relying on the so called 'statistical sufficiency'.

might be different in different observations. Anyway, for sake of generality, we shall make use of assumptions (5) and (6) only in next section.

If we think of N pairs of measurements of μ_x and μ_y, before doing the experiment we are uncertain about $4N$ quantities (all x's, all y's, all μ_x's and all μ_y's, indicated respectively as $\boldsymbol{x}$, $\boldsymbol{y}$, $\boldsymbol{\mu}_x$ and $\boldsymbol{\mu}_y$) plus the number of parameters, i.e. in total $4N+M$, that become $4N+2$ in linear fits. [But note that, due to believed deterministic relationship (3), the number of independent variables is in fact $3N+M$.] Our final goal, expressed in probabilistic terms, is to get the pdf of the parameters given the experimental information and all background knowledge:

$$\Longrightarrow f(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \quad [\rightarrow f(m,c\,|\,\boldsymbol{x},\boldsymbol{y},I) \ \text{ for linear fits}]\,.$$

Probability theory teaches us how to get the conditional pdf $f(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I)$ if we know the joint distribution $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)$. The first step consists in calculating the $2\,N+M$ variable pdf (only $N+M$ of which are independent) that describes the uncertainty of what is not precisely known, given what it is (plus all background knowledge). This is achieved by a multivariate extension of Eq. (1):

$$\begin{aligned} f(\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) &= \frac{f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)}{f(\boldsymbol{x},\boldsymbol{y}\,|\,I)} && (7)\\ &= \frac{f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)}{\int f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)\;d\boldsymbol{\mu}_x\,d\boldsymbol{\mu}_y\,d\boldsymbol{\theta}} && (8)\end{aligned}$$

Equations (7) and (8) are two different ways of writing Bayes' theorem in the case of multiple inference. Going from (7) to (8) we have 'marginalized' $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)$ over $\boldsymbol{\mu}_x$, $\boldsymbol{\mu}_y$ and $\boldsymbol{\theta}$, i.e. we used an extension of Eq. (2) to many variables. [The standard text book version of the Bayes formula differs from Eqs. (7) and (8) because the joint pdf's that appear on the r.h.s. of Eqs. (7)-(8) are usually factorized using the so called 'chain rule', i.e. an extension of Eq. (1) to many variables.]

The second step consists in marginalizing the $(2\,N+M)$-dimensional pdf over the variables we are not interested to:

$$f(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \;=\; \int f(\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I)\;d\boldsymbol{\mu}_x\,d\boldsymbol{\mu}_y \qquad (9)$$

Before doing that, we note that the denominator of the r.h.s. of Eqs. (7)-(8) is just a number, once the model and the set of observations $\{\boldsymbol{x},\boldsymbol{y}\}$ is defined, and then we can absorb it in the normalization constant. Therefore Eq. (9) can be simply rewritten as

$$f(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \;\propto\; \int f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)\;d\boldsymbol{\mu}_x\,d\boldsymbol{\mu}_y\,. \qquad (10)$$

We understand then that, essentially, we need to set up $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)$ using the pieces of information that come from our background knowledge I. This seems a horrible task, but it becomes feasible tanks to the chain rule of probability theory, that allows us to rewrite $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)$ in the following way:

$$
\begin{aligned}
f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I) &= f(\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta},I)\\
&\quad\cdot f(\boldsymbol{y}\,|\,\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta},I)\\
&\quad\cdot f(\boldsymbol{\mu}_y\,|\,\boldsymbol{\mu}_x,\boldsymbol{\theta},I)\\
&\quad\cdot f(\boldsymbol{\mu}_x\,|\,\boldsymbol{\theta},I)\\
&\quad\cdot f(\boldsymbol{\theta}\,|\,I) \qquad (11)
\end{aligned}
$$

(Obviously, among the several possible ones, we choose the factorization that matches our knowledge about of physics case.) At this point let us make the inventory of the ingredients, stressing their effective conditions and making use of independence, when it holds.

- Each observation x_i depends directly only on the corresponding true value μ_{x_i}:

  $$
  \begin{aligned}
  f(\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta},I) &= f(\boldsymbol{x}\,|\,\boldsymbol{\mu}_x,I)=\prod_i f(x_i\,|\,\mu_{x_i},I) \qquad (12)\\
  &\quad [\Longrightarrow \prod_i \mathcal{N}(\mu_{x_i},\sigma_{x_i})\ ]. \qquad (13)
  \end{aligned}
  $$

  (In square brackets is the 'routinely' used pdf.)

- Each observation y_i depends directly only on the corresponding true value μ_{y_i}:

  $$
  \begin{aligned}
  f(\boldsymbol{y}\,|\,\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta},I) &= f(\boldsymbol{y}\,|\,\boldsymbol{\mu}_y,I)=\prod_i f(y_i\,|\,\mu_{y_i},I) \qquad (14)\\
  &\quad [\Longrightarrow \prod_i \mathcal{N}(\mu_{y_i},\sigma_{y_i})\ ]. \qquad (15)
  \end{aligned}
  $$

- Each true value μ_y depends only, and in a deterministic way, on the corresponding true value μ_x and on the parameters $\boldsymbol{\theta}$. This is formally equivalent to take an infinitely sharp distribution of μ_{y_i} around $\mu_y(\mu_{x_i};\boldsymbol{\theta})$, i.e. a Dirac delta function:

  $$
  \begin{aligned}
  f(\boldsymbol{\mu}_y\,|\,\boldsymbol{\mu}_x,\boldsymbol{\theta},I) &= \prod_i \delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,] \qquad (16)\\
  &\quad [\Longrightarrow \prod_i \delta(\mu_{y_i}-m\,\mu_{x_i}-c)\ ] \qquad (17)
  \end{aligned}
  $$

- Finally, μ_{x_i} and $\boldsymbol{\theta}$ are usually independent and become the *priors* of the problem,[3] that one takes 'vague' enough, unless physical motivations suggest to do otherwise. For the μ_{x_i}

[3] Priors need to be specified for the nodes of a Bayesian network that have no *parents* (see Fig 1 and footnote 4). Priors are *logically necessary* ingredients, without which probabilistic inference is simply impossible. I understand that those who approach this kind of reasoning for the first time might be scared of this 'subjective ingredient', and because of it they might prefer methods advertised as 'objective' to which they are used, formally not depending on priors. However, if one thinks a bit deeper to the question, one realizes that behind the slogan of 'objectivity' there is much arbitrariness, of which the users are often not aware, and that might lead to seriously wrong results in critical problems. Instead, the Bayesian approach offers the logical tool to properly blend prior judgment and empirical evidence. For further comments see Ref. [2], where it is shown with theoretical arguments and many examples what is the role of priors, when they can be 'neglected' (never logically! – but almost always in routine data analysis), and even when they are so crucial that it is better to refrain from providing probabilistic conclusions.

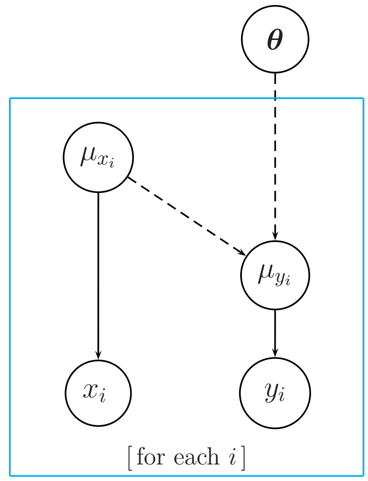

Figure 1: *Graphical representation of the model in term of a Bayesian network (see text).*

we take immediately uniform distributions over a large domain (a 'flat prior'). Instead, we leave here the expression of $f(\boldsymbol{\theta}\,|\,I)$ undefined, as a reminder for critical problems (e.g. one of the parameter is positively defined because of its physical meaning), though it can also be taken flat in routine applications with 'many' data points.

$$\begin{aligned} f(\boldsymbol{\mu}_x\,|\,\boldsymbol{\theta},I)\cdot f(\boldsymbol{\theta}\,|\,I) &= f(\boldsymbol{\mu}_x\,|\,I)\cdot f(\boldsymbol{\theta}\,|\,I) &(18)\\ &= k_x\, f(\boldsymbol{\theta}\,|\,I) &(19)\end{aligned}$$

The constant value of $f(\boldsymbol{\mu}_x\,|\,I)$, indicated here by k_x, is then in practice absorbed in the normalization constant.

In conclusion we have

$$\begin{aligned} f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I) &= \prod_i f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot\,\delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\cdot f(\mu_{x_i}\,|\,I)\cdot f(\boldsymbol{\theta}\,|\,I) &(20)\\ &= \prod_i k_{x_i}\, f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot\,\delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\cdot f(\boldsymbol{\theta}\,|\,I) &(21)\\ &\propto \prod_i f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot\,\delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\cdot f(\boldsymbol{\theta}\,|\,I)\,. &(22)\end{aligned}$$

Figure 1 provides a graphical representation of the model [or, more precisely, a graphical representation of Eq. (20)]. In this diagram the probabilistic connections are indicated by solid lines and the deterministic connections by dashed lines. These kind of networks of probabilistic and deterministic relations among uncertain quantities is known as 'Bayesian network',[4] 'belief

[4]According to *Wikipedia*[4], a Bayesian network "*is a directed graph of nodes representing variables and arcs representing dependence relations among the variables. If there is an arc from node A to another node B, then we say that A is a parent of B. If a node has a known value, it is said to be an evidence node. A node can represent*

network', 'influence network', 'causal network' and other names meaning substantially the same thing. From Eqs. (10) and (22) we get then

$$f(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \;\propto\; \left[\int\prod_i k_{x_i}\, f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot\delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\;d\boldsymbol{\mu}_x\,d\boldsymbol{\mu}_y\right]\cdot f(\boldsymbol{\theta}\,|\,I) \tag{23}$$

$$\propto\; f(\boldsymbol{x},\boldsymbol{y}\,|\,\boldsymbol{\theta},I)\cdot f(\boldsymbol{\theta}\,|\,I) = \mathcal{L}(\boldsymbol{\theta}\,;\,\boldsymbol{x},\boldsymbol{y})\cdot f(\boldsymbol{\theta}\,|\,I) \tag{24}$$

where we have factorized the unnormalized 'final' pdf into the 'likelihood'[5] $\mathcal{L}(\boldsymbol{\theta}\,;\,\boldsymbol{x},\boldsymbol{y})$ (the content of the large square bracket) and the 'prior' $f(\boldsymbol{\theta}\,|\,I)$.

We see than that, a part from the prior, the result is essentially given by the product of N terms, each of which depending on the individual pair of measurements:

$$f(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y},I) \;\propto\; \left[\prod_i \mathcal{L}_i(\boldsymbol{\theta}\,;\,x_i,y_i,I)\right]\cdot f(\boldsymbol{\theta}\,|\,I)\,, \tag{25}$$

where

$$\mathcal{L}_i(\boldsymbol{\theta}\,;\,x_i,y_i) = f(x_i,y_i\,|\,\boldsymbol{\theta},I) \;=\; k_{x_i}\int f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot\delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\;d\mu_{x_i}d\mu_{y_i} \tag{26}$$

$$=\; k_{x_i}\int f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_y(\mu_{x_i},\boldsymbol{\theta}),I)\;d\mu_{x_i} \tag{27}$$

any kind of variable, be it an observed measurement, a parameter, a latent variable, or a hypothesis. Nodes are not restricted to representing random variables; this is what is "Bayesian" about a Bayesian network." [Note: here "random variable" stands for a random variable in the frequentistic acceptation of the term ('à la von Mises' randomness) and not just as 'variable of uncertain value'.] Bayesian networks represent both a conceptual and a practical tool to tackle complex inferential problems. They have indeed renewed the interest in the field of artificial intelligence, where they are used in inferential engines, expert systems and decision makers. Browsing the web you will find plenty of applications. Here just a few references: Ref. [5] is a well known tutorial; Ref. [6] and [7] and good general books on the subject, the first of which is related to the *HUGIN* software, a lite version of it can be freely downloaded [8]; for a flash introduction to the issue, with the possibility of starting playing with Bayesian network on discrete problems JavaBayes [9] is recommended, for which I have worked also a couple of examples in [10]; for discrete and continuous variables that can be modeled with well known pdf, a good starting point is *BUGS* [11], for which I have worked out some examples concerning uncertainties in measurements [12]. BUGS stands for **B**ayesian inference **U**sing **G**ibbs **S**ampling. This means the relevant integrals we shall see later are performed by sampling, i.e. using Markov chain Monte Carlo (MCMC) methods. I do not try to introduce them here, and I suggest to look elsewhere. Good starting point can be the BUGS web page [11] and Ref. [13].

[5] Traditionally the name 'likelihood' is given to the probability of the data given the parameters, i.e. $f(\boldsymbol{x},\boldsymbol{y}\,|\,\boldsymbol{\theta},I)$, seen as a mathematical function of the parameters. Therefore the notation $\mathcal{L}(\boldsymbol{\theta}\,;\,\boldsymbol{x},\boldsymbol{y})$ [not to be confused with $f(\boldsymbol{\theta}\,|\,\boldsymbol{x},\boldsymbol{y})$!]. $f(\boldsymbol{x},\boldsymbol{y}\,|\,\boldsymbol{\theta},I)$ can be obtained marginalizing $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y\,|\,\boldsymbol{\theta},I)$, i.e. $f(\boldsymbol{x},\boldsymbol{y}\,|\,\boldsymbol{\theta},I) = \int f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y\,|\,\boldsymbol{\theta},I)\,d\boldsymbol{\mu}_x d\boldsymbol{\mu}_x$, where $f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y\,|\,\boldsymbol{\theta},I) = f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y,\boldsymbol{\theta}\,|\,I)/f(\boldsymbol{\theta}\,|\,I)$ is obtained from Eq. (20). It follows:

$$f(\boldsymbol{x},\boldsymbol{y},\boldsymbol{\mu}_x,\boldsymbol{\mu}_y\,|\,I) \;=\; \prod_i f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot\,\delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\cdot f(\mu_{x_i}\,|\,I)$$

and

$$f(\boldsymbol{x},\boldsymbol{y}\,|\,\boldsymbol{\theta},I) \;=\; \int\prod_i f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot\,\delta[\,\mu_{y_i}-\mu_y(\mu_{x_i},\boldsymbol{\theta})\,]\cdot f(\mu_{x_i}\,|\,I)\;d\mu_{x_i}d\mu_{y_i}.$$

and the constant factor k_{x_i}, irrelevant in the Bayes formula, is a reminder of the priors about μ_{x_i} (see footnote 5).

3 Linear fit with normal errors on both axes

To apply the general formulas of the previous section we only need to make explicit $\mu_{y_i}(\mu_{x_i}, \boldsymbol{\theta})$ and the error functions, and finally integrate over μ_{x_i}. In the case of linear fit with normal errors the individual contributions to the likelihoods become

$$\mathcal{L}_i(m, c\,; x_i, y_i) \;=\; k_{x_i} \int \frac{1}{\sqrt{2\pi}\,\sigma_{x_i}} \exp\left[-\frac{(x_i-\mu_{x_i})^2}{2\,\sigma_{x_i}^2}\right] \cdot \frac{1}{\sqrt{2\pi}\,\sigma_{y_i}} \exp\left[-\frac{(y_i - m\,\mu_{x_i} - c)^2}{2\,\sigma_{y_i}^2}\right] d\mu_{x_i} \tag{28}$$

$$= \; k_{x_i} \frac{1}{\sqrt{2\pi}\sqrt{\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2}} \exp\left[-\frac{(y_i - m\,x_i - c)^2}{2\,(\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2)}\right], \tag{29}$$

that, inserted into Eq. (25), finally give

$$f(m, c\,|\,\boldsymbol{x}, \boldsymbol{y}, I) \;\propto\; \prod_i \frac{1}{\sqrt{\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2}} \exp\left[-\frac{(y_i - m\,x_i - c)^2}{2\,(\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2)}\right] f(m, c\,|\,I)\,. \tag{30}$$

The effect of the error of the x-values is to have an effective standard error on the y-values that is the quadratic combination of σ_y and σ_x, the latter 'propagated' to the other coordinate via the slope m (this result can be justified heuristically by dimensional analysis).

4 Approximated solution for non-linear fits with normal errors

Linearity implies that the arguments of the exponential of the integrand in Eq. (28) contains only first and second powers of μ_{x_i}, and then the integrals has a closed solution. Though this is not true in general, the linear case teaches us how to get an approximated solution of the problem. We can take first order expansions of $\mu_y(\mu_x, \boldsymbol{\theta})$ around each x_i

$$\mu_y(\mu_{x_i}; \boldsymbol{\theta}) \;\approx\; \mu_y(x_i; \boldsymbol{\theta}) + \mu_y'(x_i; \boldsymbol{\theta}) \cdot (\mu_{x_i} - x_i)\,. \tag{31}$$

The difference $y_i - m\,\mu_{x_i} - c$ in Eq. (28), that was indeed equal to $y_i - \mu_y(\mu_{x_i}; \boldsymbol{\theta})$ in the general case, using the linear approximation becomes

$$y_i - \mu_y(x_i; \boldsymbol{\theta}) - \mu_y'(x_i; \boldsymbol{\theta}) \cdot (\mu_{x_i} - x_i) = y_i - \mu_y'(x_i; \boldsymbol{\theta}) \cdot \mu_{x_i} - [\,\mu_y(x_i; \boldsymbol{\theta}) - \mu_y'(x_i; \boldsymbol{\theta}) \cdot x_i\,]\,,$$

i.e. we have the following replacements in Eqs. (28)-(30):

$$m \;\rightarrow\; \mu_y'(x_i; \boldsymbol{\theta}) \tag{32}$$

$$c \;\rightarrow\; \mu_y(x_i; \boldsymbol{\theta}) - \mu_y'(x_i; \boldsymbol{\theta}) \cdot x_i\,. \tag{33}$$

The approximated equivalent of Eq. (30) is then

$$f(\boldsymbol{\theta}\,|\,\boldsymbol{x}, \boldsymbol{y}, I) \;\;\propto\!\approx\;\; \prod_i \frac{1}{\sqrt{\sigma_{y_i}^2 + \mu_y'^2(x_i; \boldsymbol{\theta}) \cdot \sigma_{x_i}^2}} \exp\left[-\frac{[\,y_i - \mu_y(x_i; \boldsymbol{\theta})\,]^2}{2\,[\sigma_{y_i}^2 + \mu_y'^2(x_i; \boldsymbol{\theta}) \cdot \sigma_{x_i}^2]}\right] f(\theta\,|\,I)\,, \tag{34}$$

where the unusual symbol '$\propto\!\approx$' stands for 'approximately proportional to'.

5 Extra variability of the data

As clearly stated, the previous results assume that the only sources of deviation of the measurements from the value of the physical quantities are normal errors, with known standard deviations σ_{x_i} and σ_{y_i} . Sometimes, as it is the case of the data points reported in Ref. [14], this is not the case. This means that y depends also on other, 'hidden' variables, and what we observe is the overall effects integrated over all the variability of the variables that we do not 'see'. In lack of more detailed information, the simplest modification to the model described above is to add an extra Gaussian 'noise' on one of the coordinates. For tradition and simplicity this extra noise is added to the y variable. The effect on the above result can be easily understood. Let us call σ_v the r.m.s. of this extra noise that acts normally and independently in each y point. As it is well known, the sum of Gaussian distributions is still Gaussian with an expected value and variance respectively sum of the individual expected values and variances. Therefore, the effect in the individual likelihoods (28) is to replace $\sigma^2_{y_i}$ by $\sigma^2_{y_i} + \sigma^2_v$. But we now have an extra parameter in the model, and Eq. (30) becomes

$$f(m,c,\sigma_v\,|\,\boldsymbol{x},\boldsymbol{y},I) \;\propto\; \prod_i \frac{1}{\sqrt{\sigma_v^2+\sigma_{y_i}^2+m^2\,\sigma_{x_i}^2}}\,\exp\left[-\frac{(y_i-m\,x_i-c)^2}{2\,(\sigma_v^2+\sigma_{y_i}^2+m^2\,\sigma_{x_i}^2)}\right] f(m,c,\sigma_v\,|\,I)\,. \tag{35}$$

More rigorously, this formula can be obtained from a variation of reasoning followed in the previous section.

- μ_y depends on μ_x and on the set of hidden variables $\boldsymbol{v}$:

$$\begin{aligned} \mu_y &= \mu_y^{(v)}(\mu_x,\boldsymbol{\theta},\boldsymbol{v}) && (36)\\ &= z(\mu_x,\boldsymbol{\theta}) + g(\mu_x,\boldsymbol{v}) && (37) \end{aligned}$$

  where the overall dependence $\mu_y^{(v)}(\,)$ has been split in two functions: $z(\mu_x,\boldsymbol{\theta})$, only depending on μ_x and the model parameters, corresponding to the ideal case; $g(\mu_x,\boldsymbol{v})$ describing the difference from the ideal case.

- Calling z the fictitious variable, deterministically dependent on μ_x, for a given μ_{x_i} we have the following model

$$\begin{aligned} z_i = z(\mu_{x_i},\boldsymbol{\theta})\;: &\qquad f(z_i\,|\,\mu_{x_i},\boldsymbol{\theta},I) = \delta[\,z_i - z(\mu_{x_i},\boldsymbol{\theta})\,] && (38)\\ \mu_{y_i}\;: &\qquad f(\mu_{y_i}\,|\,z_i,I) && (39) \end{aligned}$$

  where $f(\mu_{y_i}\,|\,z_i,I)$ describes our uncertainty about μ_{y_i} due to the unknown values of all other hidden variables.

- We need now to specify $f(\mu_{y_i}\,|\,z_i,I)$. As usual, in lack of better knowledge, we take a Gaussian distribution of unknown parameter σ_v, with awareness that this is just a convenient, approximated way to quantify our uncertainty.

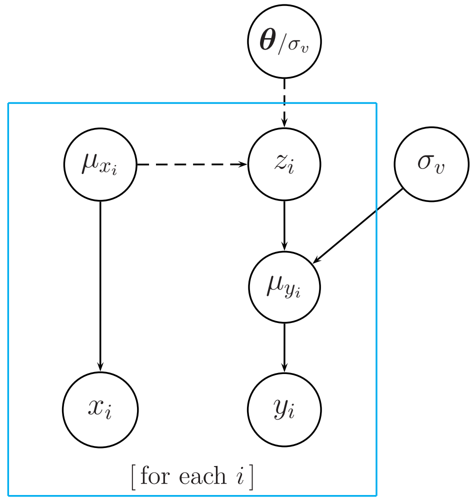

Figure 2: *Minimal modification of Fig. 1 to model the extra variability not described by the error functions. Note that $\boldsymbol{\theta}$ stands for all model parameters to be inferred, including σ_v. Instead, $\boldsymbol{\theta}/\sigma_v$ stands for all parameters apart from σ_v.*

At this point a summary of all ingredients of the model in the specific case of linear model is in order:

$$
\begin{aligned}
y_i &\sim \mathcal{N}(\mu_{y_i}, \sigma_{y_i}) && (40)\\
x_i &\sim \mathcal{N}(\mu_{x_i}, \sigma_{x_i}) && (41)\\
z_i &\leftarrow m\,\mu_{x_i} + c \qquad [\Rightarrow\ \delta(z_i - m\,\mu_{x_i} + c)\,] && (42)\\
\mu_{y_i} &\sim \mathcal{N}(z_i, \sigma_v) && (43)\\
\mu_{x_i} &\sim \mathcal{U}(-\infty, +\infty) \quad [\Rightarrow\ k_{x_i}\,] && (44)\\
m, c, \sigma_v &\Rightarrow \text{see later} \qquad [\Rightarrow \text{'uniform'}], && (45)
\end{aligned}
$$

where $\mathcal{U}(-\infty, +\infty)$ stands for a uniform distribution over a very large interval, and the symbol '$\leftarrow$' has been used to deterministically assign a value, as done in BUGS [11] (see later).

- We have now the extra parameter σ_v that we include in $\boldsymbol{\theta}$, so that M increases by 1. The new model in represented in Fig. 2, in which we have indicated by $\boldsymbol{\theta}/\sigma_v$ all parameters apart from σ_v.

- The variables of the model are now $5N + M$, and Eq. (22) becomes

$$
\begin{aligned}
f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{z}, \boldsymbol{\theta}\,|\,I) \;\propto\; & \prod_i f(x_i\,|\,\mu_{x_i}, I)\cdot f(y_i\,|\,\mu_{\mu_i}, I)\\
& \cdot f(\mu_{y_i}\,|\,z_i, I)\cdot \delta[\,z_i - z(\mu_{x_i}, \boldsymbol{\theta})\,]\cdot f(\boldsymbol{\theta}\,|\,I)\,. \qquad (46)
\end{aligned}
$$

- Consequently, Eq. (10) becomes

$$
f(\boldsymbol{\theta}\,|\,\boldsymbol{x}, \boldsymbol{y}, I) \;\propto\; \int f(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \boldsymbol{z}, \boldsymbol{\theta}\,|\,I)\; d\boldsymbol{\mu}_x\, d\boldsymbol{\mu}_y\, d\boldsymbol{z}\,. \qquad (47)
$$

- Inserting the model functions (40)-(45) in Eq. (46), after the marginalization (47) and the factorization of the result into likelihood as prior [as previously done in (24)], we get the analogues of Eqs. (26)-(28):

$$\frac{\mathcal{L}_i(\boldsymbol{\theta}\,;\,x_i,y_i)}{k_{x_i}} = \int f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot f(\mu_{y_i}\,|\,z_i,\boldsymbol{\theta},I)\cdot\delta[\,z_i - z(\mu_{x_i},\boldsymbol{\theta})\,]\;d\mu_{x_i}d\mu_{y_i}dz_i \tag{48}$$

$$= \int f(x_i\,|\,\mu_{x_i},I)\cdot f(y_i\,|\,\mu_{y_i},I)\cdot f(\mu_{y_i}\,|\,z(\mu_{x_i},\boldsymbol{\theta}),\boldsymbol{\theta},I)\;d\mu_{x_i}d\mu_{y_i} \tag{49}$$

$$= \int \frac{1}{\sqrt{2\pi}\,\sigma_{x_i}}\,\exp\left[-\frac{(x_i-\mu_{x_i})^2}{2\,\sigma_{x_i}^2}\right]\cdot\frac{1}{\sqrt{2\pi}\,\sigma_{y_i}}\,\exp\left[-\frac{(y_i-\mu_{y_i})^2}{2\,\sigma_{y_i}^2}\right] \cdot\frac{1}{\sqrt{2\pi}\,\sigma_v}\,\exp\left[-\frac{(\mu_{y_i}-m\,\mu_{x_i}-c)^2}{2\,\sigma_v^2}\right]\;d\mu_{x_i}d\mu_{y_i} \tag{50}$$

$$= \int \frac{1}{\sqrt{2\pi}\,\sigma_{x_i}}\,\exp\left[-\frac{(x_i-\mu_{x_i})^2}{2\,\sigma_{x_i}^2}\right] \cdot\frac{1}{\sqrt{2\pi}\,\sqrt{\sigma_v^2+\sigma_{y_i}^2}}\,\exp\left[-\frac{(\mu_{y_i}-m\,\mu_{x_i}-c)^2}{2\,(\sigma_v^2+\sigma_{y_i}^2)}\right]\;d\mu_{x_i} \tag{51}$$

$$= \frac{1}{\sqrt{2\pi}\,\sqrt{\sigma_v^2+\sigma_{y_i}^2+m^2\sigma_{x_i}^2}}\,\exp\left[-\frac{(\mu_{y_i}-m\,x_i-c)^2}{2\,(\sigma_v^2+\sigma_{y_i}^2+m^2\sigma_{x_i}^2)}\right] \tag{52}$$

- Inserting in Eq. (25) the expression of $\mathcal{L}_i(\boldsymbol{\theta}\,;\,x_i,y_i)$ coming from Eq. (52) we get finally Eq. (35).

6 Computational issues: normalization, fit summaries, priors and approximations

At this point it is important to understand that in Bayesian approach the full result of the inference is given by final distribution, that in our case is – we rewrite it here:

$$f(m,c,\sigma_v\,|\,\boldsymbol{x},\boldsymbol{y},I) = k\prod_i\frac{1}{\sqrt{\sigma_v^2+\sigma_{y_i}^2+m^2\,\sigma_{x_i}^2}}\,\exp\left[-\frac{(y_i-m\,x_i-c)^2}{2\,(\sigma_v^2+\sigma_{y_i}^2+m^2\,\sigma_{x_i}^2)}\right]f(m,c,\sigma_v\,|\,I)\,, \tag{53}$$

where k is 'simply' a normalization factor. (This factor is usually the most difficult thing to calculate and it is often obtained approximately by numerical methods. But this is, in principle, just a technical issue.) Once we have got k we have a full knowledge about $f(m,c,\sigma_v\,|\,\boldsymbol{x},\boldsymbol{y},I)$ and therefore about our uncertainty concerning the model parameters, the distribution of each of which can be obtained by marginalization:

$$f(m\,|\,\boldsymbol{x},\boldsymbol{y},I) = \int f(m,c,\sigma_v\,|\,\boldsymbol{x},\boldsymbol{y},I)\;dc\,d\sigma_v \tag{54}$$

$$f(c \,|\, \boldsymbol{x}, \boldsymbol{y}, I) = \int f(m, c, \sigma_v \,|\, \boldsymbol{x}, \boldsymbol{y}, I)\; dm\, d\sigma_v \tag{55}$$

$$f(\sigma_v \,|\, \boldsymbol{x}, \boldsymbol{y}, I) = \int f(m, c, \sigma_v \,|\, \boldsymbol{x}, \boldsymbol{y}, I)\; dm\, dc\,. \tag{56}$$

Similarly the joint distribution of m and c can be obtained as

$$f(m, c \,|\, \boldsymbol{x}, \boldsymbol{y}, I) = \int f(m, c, \sigma_v \,|\, \boldsymbol{x}, \boldsymbol{y}, I)\, d\sigma_v\,, \tag{57}$$

from which we can easily see that we recover Eq. (30) in the case we think the extra variability discussed in the previous section is absent. This limit case corresponds to a prior of σ_v sharply peaked around zero, i.e. $f(\sigma_v \,|\, I) = \delta(\sigma_v)$.

Other interesting limit cases are the following.

- Errors only on the y axis and no extra variability.
  Making the limit of Eq. (30) for $\sigma_{x_i} \to 0$ and neglecting irrelevant factors we get

$$f(m, c \,|\, \boldsymbol{x}, \boldsymbol{y}, I) \propto \prod_i \exp\left[-\frac{(y_i - m\,x_i - c)^2}{2\,\sigma_{y_i}^2}\right] f(m, c \,|\, I) \tag{58}$$

$$\propto \exp\left[-\frac{1}{2}\sum_i \frac{(y_i - m\,x_i - c)^2}{\sigma_{y_i}^2}\right] f(m, c \,|\, I)\,. \tag{59}$$

  This is the best known and best understood case.

- Errors only on the y axis and extra variability.
  Making the limit of Eq. (53) for $\sigma_{x_i} \to 0$

$$f(m, c, \sigma_v \,|\, \boldsymbol{x}, \boldsymbol{y}, I) \propto \prod_i \frac{1}{\sqrt{\sigma_v^2 + \sigma_{y_i}^2}} \exp\left[-\frac{(y_i - m\,x_i - c)^2}{2\,(\sigma_v^2 + \sigma_{y_i}^2)}\right] f(m, c, \sigma_v \,|\, I)\,. \tag{60}$$

- Scattering of data point around the hypothesized straight line only due to 'extra variability'.

$$f(m, c, \sigma_v \,|\, \boldsymbol{x}, \boldsymbol{y}, I) \propto \sigma_v^{-N} \prod_i \exp\left[-\frac{(y_i - m\,x_i - c)^2}{2\,\sigma_v^2}\right] f(m, c, \sigma_v \,|\, I) \tag{61}$$

$$\propto \sigma_v^{-N} \exp\left[-\frac{1}{2\,\sigma_v^2}\sum_i (y_i - m\,x_i - c)^2\right] f(m, c, \sigma_v \,|\, I)\,. \tag{62}$$

  This case corresponds to the joint determination of m, c and σ_v made by the method of the 'residuals', that can be considered a kind of approximated solution of Eq. (61), achieved by iteration. [Indeed, if there are 'enough' data points the 'best estimates' achieved by the residual method are very close to the expected values of m, c and σ_v evaluated from $f(m, c, \sigma_v \,|\, \boldsymbol{x}, \boldsymbol{y}, I)$ if we assumed a flat prior distribution for the parameters.]

Although, as it has been pointed out above, the full result of the inference is provided by the final pdf, often we do not need such a detailed description of our uncertainty, and we are only interested to provide some 'summaries'. The most interesting ones are the expected values, standard deviations and correlation coefficients, i.e. $E(m)$, $E(c)$, $E(\sigma_v)$, $\sigma(m)$, $\sigma(c)$, $\sigma(\sigma_v)$, $\rho(m,c)$, $\rho(m,\sigma_v)$ and $\rho(c,\sigma_v)$. They are evaluated from $f(m,c,\sigma_v)$ using their definitions, that are assumed to be known [hereon we often omit the conditions on which the pdf depends, and we write $f(m,c,\sigma_v)$ instead of $f(m,c,\sigma_v\,|\,\boldsymbol{x},\boldsymbol{y},I)$, and so on]. Obviously, these are not the only possible summaries. One might report in addition the mode or the median of each variable, one-dimensional or multi-dimensional probability regions [i.e. regions in the space of the parameters that are believed to contain the true value of the parameter(s) with a well defined probability level], and so on. It all depends on how standard or unusual the shape of $f(m,c,\sigma_v)$ is. I just would like to stress that the most important summaries are expected value, standard deviation and correlation coefficients, because these are the quantities that mostly matter in subsequent evaluations of uncertainty. Giving only 'most probable' values and probability intervals might bias the results of further analyzes [15].

The prior $f(m,c,\sigma_v\,|\,I)$ has been left on purpose open in the above formulas, although we have already anticipated that usually a flat prior about all parameters gives the correct result in most 'healthy' cases, characterized by a sufficient number of data points. I cannot go here through an extensive discussion about the issue of the priors, often criticized as the weak point of the Bayesian approach and that are in reality one of its points of force. I refer to more extensive discussions available elsewhere (see e.g. [2] and references therein), giving here only a couple of advices. A flat prior is in most times a good starting point (unless one uses some packages, like BUGS [11], that does not like flat prior in the range $-\infty$ to $+\infty$; in this case one can mimic it with a very broad distribution, like a Gaussian with very large σ). If the result of the inference 'does not offend your physics sensitivity', it means that, essentially, flat priors have done a good job and it is not worth fooling around with more sophisticated ones. In the specific case we are looking closer, that of Eq. (53), the most critical quantity to watch is obviously σ_v, because it is positively defined. If, starting from a flat prior (also allowing negative values), the data constrain the value of σ_v in a (positive) region far from zero, and – in practice consequently – its marginal distribution is approximatively Gaussian, it means the flat prior was a reasonable choice. Otherwise, the next-to-simple modeling of σ_v is via the step function $\theta(\sigma_v)$. A more technical choice would be a gamma distribution, with suitable parameters to 'easily' accommodate all envisaged values of σ_v.

The easiest case, that happens very often if one has 'many' data points (where 'many' might be already as few as some dozens), is that $f(m,c,\sigma_v)$ obtained starting from flat priors is approximately a multi-variate Gaussian distribution, i.e. each marginal is approximately Gaussian. In this case the expected value of each variable is close to its mode, that, since the prior was a constant, corresponds to the value for which the likelihood $\mathcal{L}(m,c,\sigma_v\,;\,\boldsymbol{x},\boldsymbol{y})$ gets its maximum. Therefore the parameter estimates derived by the maximum likelihood principle are very good approximations of the expected values of the parameters calculated directly from $f(m,c,\sigma_v)$. In a certain sense the maximum likelihood principle best estimates are recovered as a special case that holds under particular conditions (many data points and vague priors). If either condition fails, the result the formulas derived from such a principle might be incorrect. This is the reason I dislike unneeded principles of this kind, once we have a more general framework, of which the methods obtained by 'principles' are just special cases under well defined conditions.

The simple case in which $f(m, c, \sigma_v)$ is approximately multi-variate Gaussian allows also to approximately evaluate the covariance matrix of the fit parameters from the Hessian of its logarithm.[6] This is due to a well known property of the multi-variate Gaussian and it is not strictly related to flat priors. In fact it can easily proved that if the generic $f(\boldsymbol{\theta})$ is a multivariate Gaussian, then

$$(V^{-1})_{ij}(\boldsymbol{\theta}) \;=\; \left.\frac{\partial^2\varphi}{\partial\theta_i\,\partial\theta_j}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_m} \tag{63}$$

where

$$\varphi(\boldsymbol{\theta}) \;=\; -\log f(\boldsymbol{\theta})\,, \tag{64}$$

$V_{ij}(\boldsymbol{\theta})$ is the covariance matrix of the parameters and $\boldsymbol{\theta}_m$ is the value for which $f(\boldsymbol{\theta})$ gets its maximum and then $\varphi(\boldsymbol{\theta})$ its minimum.

An interesting feature of this approximated procedure is that, since it is based on the logarithm of the pdf, normalization factors are irrelevant. In particular, if the priors are flat, the relevant summaries of the inference can be obtained from the logarithm of the likelihood, stripped of all irrelevant factors (that become additive constants in the logarithm and vanish in the derivatives). Let us write down, for some cases of interest, the minus-log-likelihoods, stripped of constant terms and indicated by L, i.e. $\varphi(\boldsymbol{\theta}\,;\,\boldsymbol{x},\boldsymbol{y}) = L(\boldsymbol{\theta}\,;\,\boldsymbol{x},\boldsymbol{y}) + const.$

- Simplest case: linear fit with only known errors on the y axis [from Eq. (58)]:

$$L(m,c\,;\,\boldsymbol{x},\boldsymbol{y}) \;=\; \frac{1}{2}\sum_i \frac{(y_i - m\,x_i - c)^2}{\sigma_{y_i}^2} = \frac{1}{2}\,\chi^2(m,c\,;\,\boldsymbol{x},\boldsymbol{y})\,, \tag{65}$$

  where we recognize the famous chi-squared. Applying Eq. (63) we get then the covariance matrix of the fit parameters as

$$(V^{-1})_{m,c} \;=\; \left.\frac{1}{2}\frac{\partial^2\chi^2(m,c\,;\,\boldsymbol{x},\boldsymbol{y})}{\partial m\,\partial c}\right|_{\substack{m=m_m\\ c=c_m}} \tag{66}$$

  (See Ref. [2] for the fully developed example yielding analytic formulas for the expected values and covariance matrix of the m and c.) Note that the often used (but also often misused! [15]) '$\Delta\,\chi^2 = 1$ rule' to calculate the covariance matrix of the parameters comes from the same Gaussian approximation of the final pdf and prior insensitivity. [And, because of the factor 1/2 between Eqs. (63) and (66), there is an equivalent 'Δ minus-log-likelihood $= 1/2$' rule, applicable under the same conditions].

[6] I would like to point out that I added the formulas that follow just for the benefit of the inventory. Personally, in such low dimensional problems I find it easier to perform numerical integrations than to evaluate, obviously with the help of some software, derivatives, find minima and invert matrices, or to use the '$\Delta\,\chi^2 = 1$' or 'Δ minus-log-likelihood $= 1/2$' rules. Moreover, I think that the lazy use of computer programs solely based on some approximations produces the bad habit of taking acritically their results, even when they make no sense[15]. Nevertheless, with some reluctance and after these warnings, I give here the formulas that follows, and that the reader might know as derived from other ways, hoping he/she understands better how they can be framed in a more general scheme, and therefore when it is possible to use them.

- Errors also on the y axis:

$$L(m,c\,;\boldsymbol{x},\boldsymbol{y}) \;=\; \frac{1}{2}\sum_i \log\left(\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2\right) + \frac{1}{2}\sum_i \frac{(y_i - m\,x_i - c)^2}{\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2}\,. \tag{67}$$

  In this case expected values and covariance matrix cannot be obtained directly in closed form. Nevertheless, one can use iteratively the formulas for $\sigma_{x_i}=0$ in which the estimate of m is used to evaluate the terms $\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2$ (having the meaning of effective y-error) in the likelihood of the next iteration. Instead it is wrong to simply replace the denominator of the χ^2 of Eq. (65) with $\sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2$, because this approximation does not take into account the first term of the r.h.s. of Eq. (67) and the slope m will be underestimated (as a consequence, the intercept c will be over- or under-estimated, depending on the sign of the correlation coefficient between m and c, a sign that depends on the sign of the barycenter of the x points.)

- Dispersion on the y axis only due to σ_v [from Eq. (61)]:

$$L(m,c,\sigma_v\,;\boldsymbol{x},\boldsymbol{y}) \;=\; N\,\log\sigma_v + \frac{1}{2\,\sigma_v^2}\sum_i (y_i - m\,x_i - c)^2\,. \tag{68}$$

- The most complete case seen here [from Eq. (53)]:

$$L(m,c,\sigma_v\,;\boldsymbol{x},\boldsymbol{y}) \;=\; \frac{1}{2}\sum_i \log\left(\sigma_v^2 + \sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2\right) + \frac{1}{2}\sum_i \frac{(y_i - m\,x_i - c)^2}{\sigma_v^2 + \sigma_{y_i}^2 + m^2\,\sigma_{x_i}^2}\,. \tag{69}$$

- As the previous item, but for the general $\mu_y(\,)$ [from Eq. (34)]:

$$L(\boldsymbol{\theta},\sigma_v\,;\boldsymbol{x},\boldsymbol{y}) \approx \frac{1}{2}\sum_i \log\left[\sigma_v^2 + \sigma_{y_i}^2 + {\mu_y'}^2(x_i;\boldsymbol{\theta})\cdot\sigma_{x_i}^2\right] + \frac{1}{2}\sum_i \frac{\left[\,y_i - \mu_y(x_i;\boldsymbol{\theta})\,\right]^2}{\sigma_v^2 + \sigma_{y_i}^2 + {\mu_y'}^2(x_i;\boldsymbol{\theta})\cdot\sigma_{x_i}^2}\,. \tag{70}$$

7 From power law to linear fit

Linear fits are not only used to infer the parameters of a linear model, but also of other models that are linearized via a suitable transformation of the variables. The best known cases are the exponential law, linearized taking the log of the ordinate, and the power low, linearized taking the log of both coordinates. Linearizion is particularly important to provide a visual evidence in support of the claimed model. However, quantitative inference based on the transformed variable is not so obvious, if high accuracy in the determination of the model parameters is desired. Let us make some comments on the power law, in which both variables are log-transformed and therefore more general.

We start hypothesizing a model

$$B \;=\; \kappa\,A^{\gamma}\,, \tag{71}$$

that is linearized as

$$\log B \;=\; \gamma\, \log A + \log \kappa\,. \tag{72}$$

We identify then $\log B$ with μ_y of the linear case, $\log A$ with μ_x, γ with m and $\log\kappa$ with c. But this identification does not allows us yet to use *tout court* the formulas derived above, because each of them depends on a well defined model. Let us see where are the possible problems.

- In the simplest model a_i is normally distributed around A_i and b_i around B_i (we indicate by $\boldsymbol{a}$ and $\boldsymbol{b}$ the set of observations in the original variables). But, in general, $x_i \equiv \log a_i$ and $y_i \equiv \log b_i$ are not normally distributed around $\mu_{x_i} \equiv \log A$ and $\mu_{y_i} \equiv \log B$, respectively. They are only when the measurements are very precise, i.e. $\sigma_{a_i}/a_i \ll 1$ and $\sigma_{b_i}/b_i \ll 1$. This the case in which standard 'error propagation', based on the well known formulas base on linearization, holds.
- If the precision is not very high, i.e. σ_{a_i}/a_i and σ_{b_i}/b_i are not very small, non-linear effects in the transformations could be important (see e.g. Ref. [15]).
- When some of σ_{a_i}/a_i and σ_{b_i}/b_i approach unity it becomes important to consider the error functions and the priors about A and B with the due care. For example, very often the quantities A and B are defined positive – and if we take their logarithms, they have to be positive. This requires the model to be correctly set up in order to prevent negative values of A and B.

Further considerations would require a good knowledge of the the experimental apparatus and of the physics under study. Therefore I refrain from indicating a toy model, that could be used acritically in serious applications. Instead I encourage to draw a graphical representation of the model, as done in Figs. 1 and 2 and to make the inventory of the ingredients. Sometimes the representation in terms of Bayesian network is almost equivalent to solve the problem, thanks also to the methods developed in the past decades to calculate the relevant integrals, using e.g. Markov Chain Monte Carlo (MCMC), see e.g. Ref. [13] and references therein. In case of simple models one can even use free available software, like BUGS [11].

8 Systematic errors

Let us now consider the effect of systematic errors, i.e. errors that acts the same way on all observations of the sample, for example an uncertain offset in the instrument scale, or an uncertain scale factor. I do not want to give a complete treatment of the subjects, but focus only on how our systematic effects modify our graphical model, and give some practical rules for the simple case of linear fits. (For an introduction about systematic errors and their consistent treatment within the Bayesian approach see Ref. [2].)

For each coordinate we can introduce the fictitious quantities μ_x^S and μ_y^S that take into account the modification of μ_x and μ_y due to the systematic effect. For example, if the systematic effects only acts as an *offset*, i.e. we are uncertain about the 'true' *zero* of the instruments, ζ_x and ζ_y, we have

$$\mu_{x_i}^S \;=\; \mu_{x_i} + \zeta_x \tag{73}$$

$$\mu_{y_i}^S \;=\; \mu_{y_i} + \zeta_y\,, \tag{74}$$

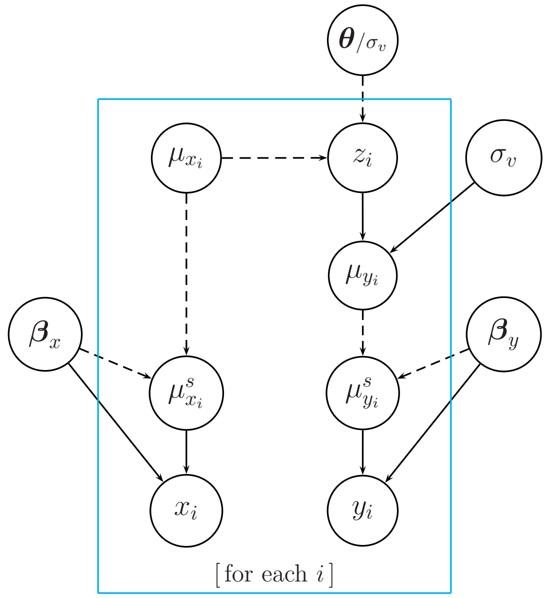

Figure 3: *Graphical model of Fig. 2 with the addition of systematic errors on both axes.*

where the true value of ζ_x are ζ_y unknown (otherwise there would be no systematic errors). We only know that their expected value is zero (otherwise we need to apply a calibration constant to the measurements) and we quantify our uncertainty with pdf's. For example, we could model them with Gaussian distributions:

$$\begin{aligned} \zeta_x &\sim \mathcal{N}(0, \sigma_{\zeta_x}) && (75) \\ \zeta_y &\sim \mathcal{N}(0, \sigma_{\zeta_y})\,. && (76) \end{aligned}$$

Anyway, for sake of generality, we leave the systematic effects in the most general form, dependent on the uncertain quantities $\boldsymbol{\beta}_x$ and $\boldsymbol{\beta}_y$ [to be clear: in the case of solely offset systematics we have $\boldsymbol{\beta}_x = \{\zeta_x\}$ $\boldsymbol{\beta}_y = \{\zeta_y\}$]. The values of $\mu^S_{x_i}$ and $\mu^S_{y_i}$ are modeled as follow

$$\begin{aligned} \mu^S_{x_i} &: \qquad \mu^S_{x_i} \leftarrow \mu^S_x(\mu_{x_i}; \boldsymbol{\beta}_x) && (77) \\ \mu^S_{y_i} &: \qquad \mu^S_{y_i} \leftarrow \mu^S_y(\mu_{y_i}; \boldsymbol{\beta}_y) && (78) \\ \boldsymbol{\beta}_x &: \qquad \boldsymbol{\beta}_x \sim f(\boldsymbol{\beta}_x \,|\, I) && (79) \\ \boldsymbol{\beta}_y &: \qquad \boldsymbol{\beta}_y \sim f(\boldsymbol{\beta}_y \,|\, I)\,. && (80) \end{aligned}$$

Figure 3 shows the graphical model containing the new ingredients. The links $\boldsymbol{\beta}_x \to x_i$ and $\boldsymbol{\beta}_y \to y_i$ are to remember that systematics could also effect the error functions. An alternative visual picture of the probabilistic model is shown in Fig. 4. Note the different symbols to indicate the different uncertain processes: the divergent arrows (in yellow, if you are reading an electronic version of the paper) indicate that, given a value of the 'parent' variable, the 'child' variable fluctuates on an event-by-event basis; the green single arrow with the question mark indicate that, given a value of the 'parent', the child will always take a fixed value, though we do not know which one.

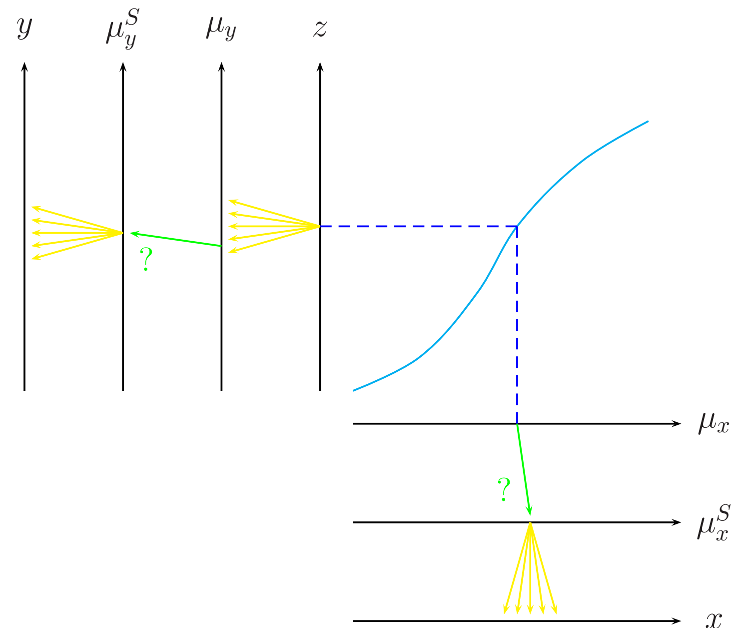

Figure 4: *A different visual representation of the probabilistic model of Fig. 3.*

Obviously, the practical implementation of complicate systematic effects in complicate fits can be quite challenging, but at least the Bayesian network provides an overall picture of the model. The simplest case is that of linear fit where only offset and scale uncertainty are present, with uncertainty modeled by a Gaussian distribution. This means that the $\boldsymbol{\beta}$'s and their uncertainty are as follows (η is the scale factor of uncertain value):

$$\boldsymbol{\beta}_x = \{\zeta_x, \eta_x\} \qquad \boldsymbol{\beta}_y = \{\zeta_y, \eta_y\} \tag{81}$$
$$\zeta_x \sim \mathcal{N}(0, \sigma_{\zeta_x}) \qquad \zeta_y \sim \mathcal{N}(0, \sigma_{\zeta_y}) \tag{82}$$
$$\eta_x \sim \mathcal{N}(1, \sigma_{\eta_x}) \qquad \eta_y \sim \mathcal{N}(1, \sigma_{\eta_y}) \tag{83}$$

In this case we can get an hint of how the uncertainty about m and c change without doing the full calculation following an heuristic approach, valid when $f(m,c)$ is approximately multivariate Gaussian and the details of which can be found in Ref. [16]. We obtain the following results, in which $\sigma(m)|_{\zeta_x}$ indicates the contribution to the uncertainty about the slope m due to uncertainty about ζ_x, $\sigma(m)|_{\eta_x}$ that due to the scale factor η_x, and so on[7]:

$$\sigma(m)|_{\zeta_x} = 0 \tag{84}$$
$$\sigma(m)|_{\zeta_y} = 0 \tag{85}$$
$$\sigma(c)|_{\zeta_x} = |m|\,\sigma_{\zeta_x} \tag{86}$$
$$\sigma(c)|_{\zeta_y} = \sigma_{\zeta_y} \tag{87}$$

[7] In Ref. [16] ζ_x is indicated by z_x, η_x by f_x, and so on.

$$\sigma(m)|_{\eta_x} = |m|\,\sigma_{\eta_x} \quad (88)$$

$$\sigma(m)|_{\eta_y} = |m|\,\sigma_{\eta_y} \quad (89)$$

$$\sigma(c)|_{\eta_x} = 0 \quad (90)$$

$$\sigma(c)|_{\eta_y} = |c|\,\sigma_{\eta_y}\,. \quad (91)$$

All contributions are then added quadratically to the so called 'statistical' ones.

9 Conclusions

The issue of fits has been approached from probability first principles, i.e. using throughout the rules of probability theory, without external *ad hoc* ingredients. It has been that the main task consists in building up the inferential model, that means in fact to properly factorize the joint probability density function of all variables of the problem. We have seen that this factorization, based on the so called chain rule of probability theory, has a very convenient graphical representation, that takes the name of Bayesian (or belief/causal/influence) network. Modeling the problem in terms of such networks not only helps to understand the problem better, but, thanks the huge amount of mathematical developments relates to them, it becomes the only way to get a (numerical) solution when problems get complicated.

We have also seen how to recover well known formulas, obtained starting from other approaches, under well defined conditions, thus indicating that other methods can be seen as approximations of the most general one, and that are therefore applicable if the conditions of validity hold.

The linear case with errors on both axis and extra variance of the data has been shown with quite some detail, giving un-normalized formulas for the pdf. In particular, going to the pretext to write this paper, we can see that Eq. (43) of Ref. [17] is not reproduced. In fact, if I understand it correctly, that equation should have the same meaning of Eq. (53) of this paper. However, Eq. (43) of Ref. [17] contains an extra factor $\sqrt{1+m^2}$ (using the notation of this paper), that it is a bit odd, for several reasons (besides the fact that I do not get it – but this could be judged a technical argument by the hurry reader). The first reason is just dimensionality: $m\,x$ is homogeneous with y and for this reason $m\,\sigma_x$ can be combined (quadratically) to σ_y, but m^2 cannot be added *tout court* to 1. The second is that if there was such a factor in Eq. (53), then one cannot reproduce Eqs. (58), (60) and (61), that one can be obtained in simpler ways (and that give rise to the likelihoods shown in Section 6, some of them rather well known). Note that the addition of a term $\sqrt{1+m^2}$ in Eq. (53) has the net effect of overestimating m, an effect that is consistent with the claim by [1] of a slope larger than that obtained by [14].[8]

[8] As a rule of thumb, since the extra variance of the data of [14] is rather important, the slope has to be very close to that obtained neglecting all σ_{x_i} and σ_{y_i} and making a very simple least square regression.

References

[1] D.E. Reichart, *Comment on astro-ph/0507588 and astro-ph/0508483*, astro-ph/0508529.

[2] G. D'Agostini, *Bayesian reasoning in data analysis: A critical introduction*, World Scientific Publishing, 2003.

[3] G. D'Agostini, *The Fermi's Bayes Theorem*, physics/0509080.

[4] `http://en.wikipedia.org/wiki/Bayesian_network`.

[5] D. Heckerman, *A Tutorial on Learning With Bayesian Networks*, MSR-TR-95-06, 1995, `http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-95-06`

[6] F.V. Jensen, *Bayesian Networks and Decision Graphs*, Springer, 2001.

[7] R.E. Neapolitan, *Learning Bayesian Networks*, Prentice Hall, 2003.

[8] HUGIN Expert, `http://www.hugin.com/Products_Services/Products/Demo/`.

[9] F.B Cozman, *"JavaBayes version 0.346 – Bayesian networks in Java"*, January 2001, `http://www-2.cs.cmu.edu/~javabayes/Home/`.

[10] `http://www.roma1.infn.it/~dagos/bn/`.

[11] BUGS, *The BUGS Project - Bayesian inference Using Gibbs Sampling*, `http://www.mrc-bsu.cam.ac.uk/bugs/welcome.shtml`.

[12] `http://www.roma1.infn.it/~dagos/BUGS/`.

[13] C. Andrieu at al., *An Introduction to MCMC for machine learning*, Machine Learning 2002 (preprint `http://www.cs.ubc.ca/~nando/papers/mlintro.ps` and mlintro.pdf).

[14] C. Guidorzi, *Testing the GRB Variability/Peak Luminosity correlation using the pseudo-redshifts of a large sample of BATSE GRBs*, astro-ph/0508483.

[15] G. D'Agostini, *Asymmetric Uncertainties: Sources, Treatment and Potential Dangers*, physics/0403086.

[16] G. D'Agostini, *Errori e incertezze di misura. Rassegna critica e proposte per l'insegnamento*, Nota Interna 1094, 1998, pp. 69-70, `http://www.roma1.infn.it/~dagos/perfezionamento.ps` (and pdf), `http://www.roma1.infn.it/~dagos/perfezionamento/node57.html`.

[17] D.E. Reichart, *Dust extinction curves and LYα forest flux deficits for use in modeling GRB afterglows and all other extragalactic point sources*, ApJ **553** (2001) 235 (preprint astro-ph/9912368).