Bayes in five days

V. Dose

Centre for Interdisciplinary Plasma Science,

Max-Planck-Institut für Plasmaphysik, EURATOM Association,

Boltzmannstr. 2, 85748 Garching, Germany

(Dated: November 19, 2002)

Lecture notes from a ten hour tutorial on Bayesian analysis given at the International Max-Planck Research School on bounded plasmas

Greifswald May 14-18, 2002

Contents

I. BAYESIAN PRINCIPLES

R.T.Cox (Am. J. Phys. 14 (1946)1) has shown from the requirement of logical consistency that a probability theory named after Reverend Thomas Bayes can be derived from exactly two rules. The first is the product rule and tells how to break down a probability or probability density $p(D, H|I)$ into simpler functions

$$p(H, D|I) = p(H|I) \cdot p(D|H, I) \tag{1}$$

The first term on the right hand side does not require further explanation. The second factor is the conditional probability of D given that H and I are true. Strictly, the left hand side of (1) is also a conditional probability with condition I which summarizes all the background of a person which allowed him to specify $p(H, D|I)$. This function is symmetric in the variables H and D and consequently there exists an alternative expansion.

$$p(H, D|I) = p(D|I) \cdot p(H|D, I) \tag{2}$$

The combination (1+2) of the two expansions results in Bayes' theorem

$$p(H|D, I) = p(H|I) \cdot p(D|H, I)/p(D|I) \tag{3}$$

Let us identify H with "hypothesis" and D with "data". $p(H|I)$ is called the prior probability of H, it specifies expert knowledge on H before an experiment designed to provide data D is performed. $p(D|H, I)$ is the sampling distribution of the data "D" when regarded as a function of D. We call it the likelihood function when regarded as a function of H given a fixed set of data D. A warning is here in order. There is no symmetry between variables and conditions, therefore

$$\int p(D|H, I)dD = 1 \neq \int p(D|H, I)dH \tag{4}$$

$p(D|H, I)$ is our expectation of possible data sets assuming that we know the physics ($= H$). $p(D|H, I)$ is therefore the function which encodes our model of the data. We postpone a discussion on the meaning of $p(D|I)$ in (3). It is a constant in the present context which we will show to ensure normalization.

Bayes theorem allows for an inverse conclusion. Given that our model H allows to fit the data, the product

$$p(H|I) \cdot p(D|H, I) \tag{5}$$

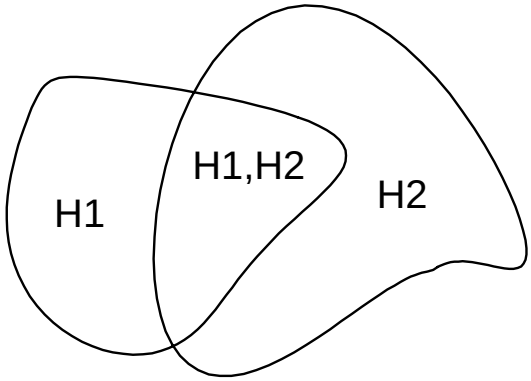

FIG. 1:

allows us to assign the probability $p(H|D,I)$ of the truth of our model. This was a major progress in medieval philosophy, since at Galilei's time the conclusion inversion "Animals, which move, have extremities and muscles. The earth has neither extremities nor muscles, consequently it does not move" was considered to be conclusive proof against his observations.

We now proceed to the second rule, the sum rule. Given two hypotheses H_1 and H_2 it says

$$p(H_1 + H_2|I) = p(H_1|I) + p(H_2|I) - p(H_1, H_2|I) \tag{6}$$

The plus sign "$H_1 + H_2$" should be read as an "or" while the comma "H_1, H_2" should be read as a logical "and". The meaning of the sum rules is best explained with the reference to Fig. 1. The probability that H_1 or H_2 are true is equal to the probability that either H_1 or H_2 are true minus the probability that H_1 and H_2 are simultaneously true. An important special case arises if the H_i are exhaustive and mutually exclusive. Mutually exclusive means no overlap diagrammatically. Exhaustive means that they cover all conceivable possibilities. We then have

$$p(H_1 + H_2 + ...|I) = p(\sum_i H_i|I) = 1 \tag{7}$$

Let us then consider the function $p(D, \sum_i H_i|I)$. First we apply the product rule to obtain

$$p(D, \sum_i H_i|I) = p(D|I) \cdot p(\sum_i H_i|D, I) = p(D|I) \tag{8}$$

The last equality in (8) holds because of (7) since regardless of what the data D are the

H-space is exhaustive.

$$p(\sum_i H_i|D,I) = p(\sum_i H_i|I) = 1 \tag{9}$$

In a case where a conditional probability $p(\sum_i H_i|D,I)$ is independent of the conditional variables we call it logically independent of the respective conditions. We shall make ample use of logical independence in the following.

Next we apply the sum rule (7) to the left hand side of (8) in the reverse direction to obtain

$$p(D,\sum_i H_i|I) = \sum_i p(D,H_i|I) \tag{10}$$

By comparison with the right hand side of (8) we obtain the important marginalization rule

$$p(D|I) = \sum_i p(D,H_i|I) \tag{11}$$

As a first example for the application of sum and product rule we consider an urn with w(hite) and b(lue) balls with masses m, M. The following table describes the contents of the urn. If we ask for the probability to draw a white ball regardless of its mass, the answer is

α, β	#	$p(\alpha,\beta)$
w, m	100	0.1
w, M	200	0.2
b, m	300	0.3
b, M	400	0.4

provided by the sum rule

$$p(w) = p(w,m) + p(w,M) \tag{12}$$

Inserting the numbers from the table yields $p(w) = 0.3$. Next we ask for the probability of drawing a ball with mass M given that it is white. We apply the product rule

$$p(M,w) = p(w)\cdot p(M|w) \rightarrow p(M|w) = p(M,w)/p(w) \tag{13}$$

Inserting numbers from the table and (12) yields $p(M|w) = 2/3$. It is instructive to convince oneself at this point that $p(M|w) \neq p(w|M)$. Expand the left hand side of (13) alternatively as

$$p(M,w) = p(M)\cdot p(w|M) \rightarrow p(w|M) = p(M,w)/p(M) \tag{14}$$

which yields $p(w|M) = 1/3$. This comparison highlights the asymmetry between arguments and conditions in a conditional probability.

As a second application of the sum and product rule we consider a problem whose solution is not immediately obvious. A show master acts on a stage with three doors. Behind one of the doors is hidden a price, e.g. a car. The candidate has to choose one of the three doors which remains, however, closed. The show master opens another door, of course one which leaves the prize hidden. The candidate has now to make his second choice between two doors to open. Should he stay with his first choice or should he change his mind? Does it matter at all wether he changes or not? To find the answer we need the probability $p(W|S, I)$ where "W" now stands for "Win" and "S" for "Strategy" and the rules of the game are summarized in "I". With the definitions

W: win

S: choose new door

σ: a nuisance variable with $\sigma = 1$: first choice correct and $\sigma = 0$: first choice wrong. We obtain $p(W|S, I)$ by application of the sum rule

$$p(W|S, I) = \sum_{\sigma=0}^{1} p(W, \sigma|S, I) \tag{15}$$

Application of the product rule to the sum terms transforms this into

$$p(W|S, I) = \sum_{\sigma=0}^{1} p(\sigma|S, I)p(W|\sigma, S, I) \tag{16}$$

The probability $p(\sigma|S, I)$ is logically independent of S and takes on the values $p(\sigma = 1|I) = 1/3$ and $p(\sigma = 0|I) = 2/3$. The second term $p(W|\sigma, S, I)$ yields 1 for $\sigma = 0$, because if the first choice was wrong the rules of the game dictate that the show master opens the other "no win" door. Correspondingly $p(W|\sigma = 1, S, I) = 0$ because if the first choice was correct, a subsequent change cannot lead to win. Therefore

$$\begin{aligned} p(W|S, I) = &\underbrace{p(\sigma = 0|I)}_{2/3} \cdot \underbrace{p(W|\sigma = 0, S, I)}_{1} \\ &+ \underbrace{p(\sigma = 1|I)}_{1/3} \cdot \underbrace{p(W|\sigma = 1, S, I)}_{0} = \frac{2}{3} \end{aligned} \tag{17}$$

The strategy S to choose a new door is therefore advantageous since the probability to win on application of this strategy is 2/3.

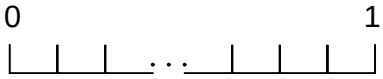

FIG. 2: Range and possible values of a parameter provide an exhaustive and mutually exclusive set of hypotheses.

A. Probability distributions

In the previous two examples the variable H took on only two discrete values. A transition to the continuous case is obtained if we allow the numerical values x_i associated with H to cover a certain range, e.g. $0 \leq x \leq 1$. Let x_1 be the hypothesis that $0 \leq H < 0.1$, x_2 the hypothesis that $0.2 \leq H < 0.3$, etc. These alternative hypotheses are exhaustive since they cover the whole allowed range $[0, 1]$. They are also exclusive since a measurement on H cannot provide two values at the same time. This means

$$p(x_1|I) + p(x_2|I) + ... = \sum_i p(x_i|I) = 1 \rightarrow \int p(x|I)dx = 1 \tag{18}$$

Now consider a second variable y

$$\begin{array}{rcl} p(y, x_1|I) & = & p(y|I) \cdot p(x_1|y, I) \\ \vdots & & \vdots \\ p(y, x_N|I) & = & p(y|I) \cdot p(x_N|y, I) \\ \hline \sum_k p(y, x_k|I) & = & p(y|I) \cdot \sum_k p(x_k|y, I) \end{array} \tag{19}$$

The second factor on the right hand side is in analogy to (8) equal to one which yields

$$\int p(y, x|I)dx = p(y|I) = \int p(x|I) \cdot p(y|x, I)dx \tag{20}$$

(20) is the extremely important marginalization rule for the case of continuous probability densities. We shall use it repeatedly in the sequel. Before we discuss an example for multivariate, conditional, and marginal distributions we give a short account on one-dimensional Gaussian integrals.

B. Gaussian integrals

A frequent analytical form, either exact or as an approximation, for a probability distribution is a Gaussian function G. We consider

$$G(x|A,B,C) = \frac{\exp\{-\frac{A}{2}x^2 + Bx - \frac{C}{2}\}}{Z} \tag{21}$$

where Z ensures normalization. Explicitly

$$Z = \int_{-\infty}^{\infty} \exp\{-\frac{A}{2}x^2 + Bx - \frac{C}{2}\}dx \tag{22}$$

First we transform the argument of the exponential into a complete square plus a residue

$$\frac{A}{2}x^2 - Bx + \frac{C}{2} = \frac{Q}{2}(x - x_o)^2 + R \tag{23}$$

Comparing coefficients yields

$$\begin{aligned} Q = A, \qquad Qx_0 = B, \qquad & R = \tfrac{1}{2}(C - Qx_0^2) \\ x_0 \;=\; B/A, \quad & R = \tfrac{1}{2}(C - B^2/A) \end{aligned} \tag{24}$$

The integral becomes

$$Z = \exp\{-\frac{1}{2}(C - B^2/A)\} \cdot \int_{-\infty}^{\infty} \exp\{-\frac{A}{2}(x - x_0)^2\}dx \tag{25}$$

The substitution $\xi = (x - x_0) \cdot \sqrt{A/2}$ transforms this into

$$Z = \sqrt{\frac{2}{A}} \exp\{-\frac{1}{2}(C - B^2/A)\} \int_{-\infty}^{\infty} \exp\{-\xi^2\}d\xi \tag{26}$$

The remaining integral $I = \int_{-\infty}^{\infty} \exp\{-\xi^2\}d\xi$ can be found easily via

$$I^2 = \int_{-\infty}^{\infty} d\xi \int_{-\infty}^{\infty} d\eta e^{-(\xi^2+\eta^2)} = 2\pi \int_{0}^{\infty} \rho d\rho e^{-\rho^2} = \pi \tag{27}$$

so that the normalization integral Z becomes

$$Z = \sqrt{\frac{2\pi}{A}} \exp\{-\frac{1}{2}(C - B^2/A)\} \tag{28}$$

Later on we shall wish to evaluate expectation values for x and x^2

$$< x >= \frac{1}{Z} \int_{-\infty}^{\infty} x \cdot \exp\{-\frac{A}{2}x^2 + Bx - \frac{C}{2}\}dx \tag{29}$$

$$< x^2 >= \frac{1}{Z} \int_{-\infty}^{\infty} x^2 \exp\{-\frac{A}{2}x^2 + Bx - \frac{C}{2}\}dx \tag{30}$$

(29) can be expressed as

$$< x >= \frac{1}{Z} \cdot \frac{\partial}{\partial B} \int_{-\infty}^{\infty} dx \exp\{-\frac{A}{2}x^2 + Bx - \frac{C}{2}\} = \frac{1}{Z}\frac{\partial Z}{\partial B} = \frac{\partial}{\partial B}\log Z \tag{31}$$

Similarly we obtain for (30)

$$< x^2 >= \frac{1}{Z}(-2\frac{\partial}{\partial A}) \cdot \int_{-\infty}^{\infty} dx \exp\{-\frac{A}{2}x^2 + Bx - \frac{C}{2}\} = -2\frac{\partial}{\partial A}\log Z \tag{32}$$

Using (28) we obtain explicitly

$$< x >= B/A, \qquad < x^2 >= 1/A + B^2/A^2 \tag{33}$$

More interesting than the expectation of x^2 is its variation with respect to the mean value $< x >$. This is called the variance

$$\begin{aligned} var(x) &=< (x- < x >)^2 > \\ var(x) &=< x^2 > -2 < x < x >> + < x >^2 =< x^2 > - < x >^2 \end{aligned} \tag{34}$$

with the previous results (33) the variance becomes

$$var(x) =< \Delta x^2 >= 1/A \tag{35}$$

C. Product rule and marginalization

Let us consider a probability distribution of two variables x, y. $p(x, y)$ is positive semidefinite and constitutes a bubble above the $(x-y)$ plane. As an example we consider a Gaussian functional relationship

$$p(x, y) = \exp\{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2} - 2a\frac{xy}{2\sigma_x\sigma_y}\}/Z \tag{36}$$

Using the product rule $p(x,y)$ may be written as

$$p(x,y) = p(x) \cdot p(y|x) \tag{37}$$

In order to understand the three distributions in (37) better we shall calculate the right hand side from the form (36). First we calculate p(x) using the marginalization rule (20)

$$p(x) = \frac{1}{Z} \int_{-\infty}^{\infty} dy \exp\{-\frac{x^2}{2\sigma_x^2} - \frac{y^2}{2\sigma_y^2} - \frac{axy}{\sigma_x \sigma_y}\} \tag{38}$$

The integral is of the type (22) with $A = 1/\sigma_y^2 \quad C = x^2/\sigma_x^2$ and $B = ax/(\sigma_x \sigma_y)$. Our standard formula (28) leads then to

$$p(x) = \frac{1}{Z} \cdot \sigma_y \sqrt{2\pi} \exp\{-\frac{x^2}{2\sigma_x^2}(1-a^2)\} \tag{39}$$

from which we can read using (28) immediately the normalization

$$\int p(x)dx = \frac{1}{Z}\sigma_y\sqrt{2\pi} \cdot \sigma_x\sqrt{2\pi}/\sqrt{1-a^2} \tag{40}$$

hence

$$p(x) = \sqrt{\frac{1-a^2}{2\pi\sigma_x^2}} \exp\{-\frac{x^2}{2\sigma_x^2}(1-a^2)\} \tag{41}$$

Knowing $p(x)$ and $p(x,y)$ we obtain the conditional distribution $p(y|x)$ from

$$p(y|x) = p(x,y)/p(x) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma_y^2}(y + ax \cdot \frac{\sigma_y}{\sigma_x})^2\} \tag{42}$$

$p(y|x)$ is a series of one dimensional Gaussian functions in y shifted by $-ax\sigma_y/\sigma_x$. Fig.3 depicts the three distributions.

D. Likelihood, prior, evidence and posterior

A proper understanding of the probabilities which enter Bayes' theorem is of crucial importance for a successful application of the theory. We shall now explain the various terms by reference to a particularly simple example. Consider a set of measurements d_i on one and the same quantity μ (think for example of the quantum Hall constant)

$$\begin{aligned} d_1 = \mu \pm \varepsilon_1 \qquad & d_1 - \mu = \pm\varepsilon_1 \\ \vdots \qquad\qquad & \vdots \\ d_N = \mu \pm \varepsilon_N \qquad & d_N - \mu = \pm\varepsilon_N \end{aligned} \tag{43}$$

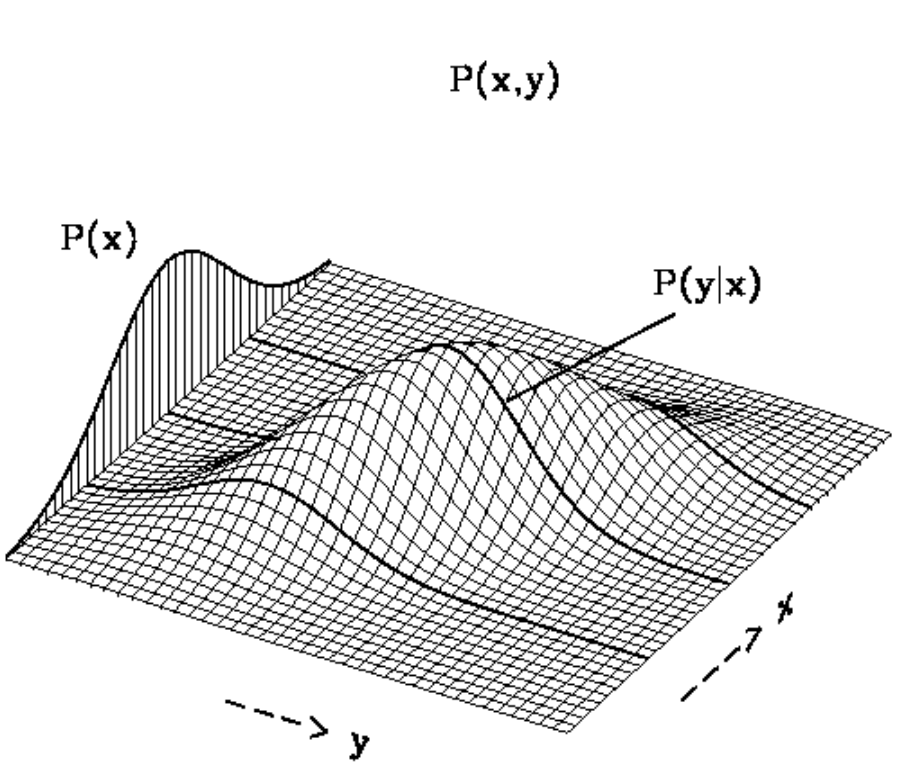

FIG. 3: The two variable distribution $p(x, y)$ and the associated marginal $p(x)$ and conditional $p(y|x)$ distributions.

We now make the strong assumption that the measurement errors ε_i are Gaussian distributed with zero mean and variance σ^2. We shall show in the next chapter that the assumption "Gaussian distributed" can be relaxed in the framework of the principle of maximum entropy leading to the same result, namely

$$p(\varepsilon|I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{\varepsilon^2}{2\sigma^2}\} = p(d|\mu, \sigma, I), \quad p(d|\mu, \sigma, I) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(d-\mu)^2\} \quad (44)$$

For N measurements we have to consider

$$p(d_1, d_2, ...d_N|\mu, \sigma, I) = p(d_1|\mu, \sigma) \cdot p(d_2, ...d_N|d_1, \mu, \sigma, I) \quad (45)$$

where we have used the product rule to expand the left hand side of (45). Now, if the measurements $d_2...d_N$ are independent of the outcome of experiment 1 then the second factor in (45) is (logically) independent of d_1.

By repeated application of the product rule and the argument of logical independence we obtain

$$p(\vec{d}|\mu, \sigma, I) = (\frac{1}{\sigma\sqrt{2\pi}})^N \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\} \quad (46)$$

$p(\vec{d}|\mu,\sigma,I)$ when regarded as a function of μ is our likelihood function. Since we shall σ assume to be known for the moment the only quantity which we want to infer from the likelihood is the numerical value of μ. To this end we need the distribution $p(\mu|\vec{d},\sigma,I)$ which we obtain from Bayes' theorem

$$p(\mu|\vec{d},\sigma,I) = p(\mu|\sigma,I)\cdot p(\vec{d}|\mu,\sigma,I)/p(\vec{d}|\sigma,I) \tag{47}$$

$p(\mu|\sigma,I)$ is the prior probability which we associate with μ. Since $p(\mu|\sigma,I)$ is our knowledge before the experiment to obtain $\vec{d}$ is carried out and since this experiment specifies σ we conclude that $p(\mu|\sigma,I) = p(\mu|I)$ is independent of σ. The prior probability of μ has now to be specified.

Different persons will, according to their different expertise come to different conclusions here. We shall assume that earlier measurements have yielded a value μ_0 of μ with standard deviation S. We shall show in the next chapter that this state of knowledge specifies $p(\mu|I)$ as

$$p(\mu|\mu_0,S,I) = \frac{1}{S\sqrt{2\pi}}\exp\{-\frac{1}{2S^2}(\mu-\mu_0)^2\} \tag{48}$$

The final quantity we need in Bayes' theorem to completely specify the posterior distribution $p(\mu|\vec{d},\sigma,I)$ is the so called evidence $p(\vec{d}|\sigma,I)$. This distribution is not independent of prior and likelihood. We rather find that

$$p(\vec{d}|\sigma,I) = \int p(\vec{d},\mu|\sigma,I)d\mu = \int p(\mu|\sigma,I)\cdot p(\vec{d}|\mu,\sigma,I)d\mu \tag{49}$$

It therefore serves to normalize the right hand side of (47) and thereby provides a normalized posterior. We shall calculate $p(\vec{d}|\sigma,I)$ first using the prior (48) and the likelihood (46)

$$p(\vec{d}|\sigma,I) = \frac{1}{S\sqrt{2\pi}}\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \int d\mu \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2 - \frac{1}{2S^2}(\mu-\mu_0)^2\right\} \tag{50}$$

We introduce the notation $\sum d_i = N\overline{d}$ and $\sum d_i^2 = N\overline{d^2}$ and order the exponent in (50) in decreasing powers of μ

$$\begin{aligned} -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu) - \frac{1}{2S^2}(\mu-\mu_0)^2 &= -\frac{N}{2\sigma^2}\left(\overline{d^2} - 2\mu\overline{d} + \mu^2\right) - \frac{1}{2S^2}\left(\mu^2 - 2\mu\mu_0 + \mu_0^2\right) \\ &= -\frac{\mu^2}{2}\underbrace{\left\{\frac{N}{\sigma^2}+\frac{1}{S^2}\right\}}_{A} + \mu\underbrace{\left\{\frac{N\overline{d}}{\sigma^2}+\frac{\mu_0}{S^2}\right\}}_{B} - \frac{1}{2}\underbrace{\left\{\frac{N\overline{d^2}}{\sigma^2}+\frac{\mu_0^2}{S^2}\right\}}_{C} \end{aligned} \tag{51}$$

The notation A, B, C corresponds to the notation adopted for the evaluation of Gaussian integrals (22-35) and allows us to derive immediately mean and variance of μ as

$$\begin{aligned} < \mu > &= \frac{B}{A} = \frac{N\overline{d}/\sigma^2 + \mu_0/S^2}{N/\sigma^2 + 1/S^2} \\ < \Delta\mu^2 > &= \frac{1}{A} = \frac{1}{N/\sigma^2 + 1/S^2} = \frac{S^2\sigma^2}{NS^2 + \sigma^2} \end{aligned} \tag{52}$$

It is interesting to investigate the large N limit of the expressions (52). For large N the expectation value $< \mu >$ becomes $< \mu > \rightarrow \overline{d}$ independent of the prior knowledge. The same happens for the variance which becomes $< \Delta\mu^2 > \rightarrow \sigma^2/N$. For moderate and small N we must refine the discussion. Consider the ratio R

$$R = \frac{N\overline{d}}{\sigma^2} \cdot \frac{S^2}{\mu_0} = N\frac{\overline{d}}{\mu_0} \cdot \frac{S^2}{\sigma^2} \tag{53}$$

If our prior knowledge happens to estimate the outcome of the experiment correctly, e.g. $\overline{d} \approx \mu_0$ and further if the prior estimate is of similar precision as the measurement $S \approx \sigma$ then the amount of prior knowledge which enters the posterior estimate corresponds to exactly one data point d(!). If the prior knowledge is not that precise e.g. S larger than σ, then the prior information corresponds only to a fraction of a measurement. In all cases since

$$1/\left(N/\sigma^2 + 1/S^2\right) < 1/\left(N/\sigma^2\right) \tag{54}$$

the informative prior (48) decreases the variance of the posterior, but the amount declines rapidly as the number N of data increases.

Finally we investigate $p(\vec{d}|\sigma, \mu_0, S, I)$. From (51) we have

$$p(\vec{d}|\sigma, \mu_0, S, I) = \frac{1}{S\sqrt{2\pi}} \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \sqrt{\frac{2\pi}{A}} \exp\left\{-\frac{1}{2}(C - B^2/A)\right\} \tag{55}$$

which we now want to investigate as a function of μ_0. Since only the coefficients B and C are μ_0 dependent it is sufficient to look at

$$p(\vec{d}|\sigma, \mu_0, S, I) \sim \exp\left\{-\frac{1}{2}\left[\frac{N\overline{d^2}}{\sigma^2} + \frac{\mu_0^2}{S^2} - \frac{(\frac{N\overline{d}}{\sigma^2} + \frac{\mu_0}{S^2})^2}{N/\sigma^2 + 1/S^2}\right]\right\} \tag{56}$$

For a given data set with $\overline{d}$ and $\overline{d^2}$, a data summary which is called a sufficient statistic, the function (56) tends to zero as $\mu_0 \rightarrow \pm\infty$. Moreover, since it is positive everywhere it must

have a maximum at finite μ_0 which we find from

$$\frac{d}{d\mu_0}\left\{\frac{N\overline{d^2}}{\sigma^2}+\frac{{\mu_0}^2}{S^2}-\left(\frac{N\overline{d}}{\sigma^2}+\frac{\mu_0}{S^2}\right)^2\frac{\sigma^2S^2}{\sigma^2+NS^2}\right\}=0 \tag{57}$$

to be

$$\hat{\mu}_0=\overline{d} \tag{58}$$

The evidence has a maximum at $\hat{\mu}_0=\overline{d}$, that is exactly when we happen to dispose of prior information equal to what the data tell. Another frequently used expression for "evidence" is (now understandable) "prior predictive value".

E. Iteration of the data information

Many people, who first learn about Bayes' theorem propose the cunning idea to use it iteratively. This means, when we have obtained a posterior distribution from Bayes' theorem would it not be a good idea to use this posterior as a prior for a reanalysis of the data? The kind of question anticipates the answer: no(!). We shall demonstrate what happens using the likelihood (46) as an example

$$p(\vec{d}|\mu,\sigma,I)=\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\right\} \tag{59}$$

Our aim is to estimate μ. Bayes' theorem yields

$$p(\mu|\vec{d},\sigma,I)=p(\mu|I)\cdot p(\vec{d}|\mu,\sigma,I)/p(\vec{d}|\sigma,I) \tag{60}$$

In order to keep things simple we choose a flat prior for μ

$$p(\mu|I)=\frac{1}{\Delta},\quad \Delta=\mu_{max}-\mu_{min},\quad \mu_{min}<\mu<\mu_{max} \tag{61}$$

with Δ sufficiently large such that the likelihood is "essentially" zero outside the range of $p(\mu|I)$. The posterior distribution is then

$$p(\mu|\vec{d},\sigma,I)=\frac{1}{\Delta}\cdot\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\right\}/p(\vec{d}|\sigma,I) \tag{62}$$

We shall only ask for the maximum and width of this distribution as a function of μ. We may then go over to proportionalities and delete all terms which do not depend on μ hence

$$p^{(1)}(\mu|\vec{d},\sigma)\sim\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\right\} \tag{63}$$

Now applying Bayes' theorem once more with (63) as a prior and (59) as likelihood yields the posterior $p^{(2)}$

$$p^{(2)} \sim p^{(1)}(\mu|\vec{d},\sigma,I)\cdot p(\vec{d}|\mu,\sigma,I) \sim \exp\left\{-\frac{2}{2\sigma^2}\sum_{i=1}^{N}(d_i-\mu)^2\right\} \tag{64}$$

$p^{(2)}$ has the same maximum as $p^{(1)}$ namely $\hat{\mu} = \overline{d}$. Note however that the variance has decreased by a factor of 2. M iterations of the above kind would still yield a maximum of $\hat{\mu} = \overline{d}$ but a variance smaller by a factor of M compared to (63). Since we can make M arbitrarily large we can generate a result independent of the order of iteration but with arbitrarily high precision which is obviously a complete nonsense.

F. Posterior characterization

The full answer of a Bayesian analysis is the posterior distribution. This can easily be visualized in two and three dimensions. Apart from the fact that this leaves the question how to proceed in higher dimensions it is quite useful to summarize the posterior in terms of a few numbers. We have already used the mean (29) and the variance (34) to obtain a point estimate of a distribution and its confidence limits. In case of a Gaussian distribution

$$G(x|x_0,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}(x-x_0)^2\right\} \tag{65}$$

these are

$$< x >= x_0, \qquad \sqrt{< \Delta x^2 >} = \pm\sigma \tag{66}$$

The situation is depicted in Fig. 4 which shows a Gaussian function centered at x_0 and the positions $x_0 \pm \sigma$. We note that the area in each of the wings is 0.1586.

In case of a Gaussian function we may have equally well chosen the maximum and the curvature at the maximum. The curvature at the maximum, approximated by the second derivative is

$$f''(x_0) = -\frac{1}{\sigma^2} \tag{67}$$

So things are simple and straight forward for a Gaussian. A frequently chosen procedure for non Gaussian posterior distributions is the Gaussian approximation. Let

$$p(x|\vec{d},\sigma,I) = \exp\left\{-\phi(x)\right\}/Z \tag{68}$$

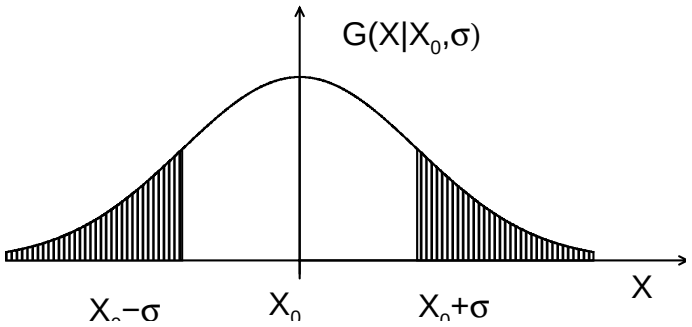

FIG. 4: Approximately 68% of the mass of a Gaussian probability distribution is concentrated in the region $(x_0 - \sigma) \leq x \leq (x_0 + \sigma)$

Now expand the argument of the exponential in a Taylor series up to second order in x around the maximum x_m of $p(x|\vec{d}, \sigma, I)$

$$\phi(x) = \phi(x_m) + \frac{\partial \phi}{\partial x}|_{x_m}(x - x_m) + \frac{1}{2} \quad \frac{\partial^2 \phi}{\partial x^2}|_{x_m}(x - x_m)^2 \tag{69}$$

At the maximum of ϕ the first derivative term vanishes and the coefficient of $(x - x_m)^2$ is equal to

$$\frac{\partial^2 \phi}{\partial x^2}|_{x_m} = \frac{1}{\sigma^2} \tag{70}$$

The Gaussian approximation, though meaningful in many cases, may not be applicable. This is true if the original function is strongly asymmetric, multimodal, or has no well defined moments at all. A simple case, where the Gaussian approximation would be misleading, is the Lorentzian

$$\mathcal{L}(x|\gamma) = \frac{\gamma}{2\pi} \frac{1}{x^2 + \gamma^2/4} \tag{71}$$

Both the first and the second moments are not defined on this function since the integral I_n

$$I_n = \frac{\gamma}{2\pi} \int_{-\infty}^{Z} \quad \frac{x^n dx}{x^2 + \gamma^2/4} \tag{72}$$

diverges already for $n = 1$. A useful alternative in this case is the median $\hat{x}$. It is defined via

$$\int_{-\infty}^{\hat{x}} p(x|I)dx = \frac{1}{2} \cdot \int_{-\infty}^{\infty} p(x|I)dx \tag{73}$$

There are two advantages for this definition. The median does exist for every normalizable probability distribution and is also meaningful for multimodal and asymmetric distributions.

But apart from the point estimate $\hat{x}$ we want usually also some measure for confidence limits. There is no generally agreed solution to this operation. One possible definition would be to choose those values x_{lo} and x_{hi} for which

$$\int_{-\infty}^{x_{lo}} p(x|I)dx = \int_{x_{hi}}^{\infty} p(x|I)dx = 0.1586... \tag{74}$$

in analogy to the Gaussian case (see Fig.4). For a Lorentzian this recipe yields

$$\frac{\gamma}{2\pi}\int_{-\infty}^{x_{lo}} \frac{dx}{x^2+\gamma^2/4} = \frac{1}{\pi}\left\{\frac{\pi}{2} - arctg\frac{2x_{lo}}{\gamma}\right\} = 0.16 \tag{75}$$

and after solving for x_{lo}, $x_{lo} = -0.92\gamma$ corresponding to roughly twice the half width while the Gaussian approximation yields $x_{lo} = \sigma = -\gamma/\sqrt{2}$. (74) allows of course to define asymmetric confidence limits. In summary we point out that the full answer to any problem is the complete posterior. Caution has to be observed when characterizing it by just a few numbers.

G. Sequential or one step estimation

Suppose we have a set of data $\vec{d}$ which suggests some obvious decomposition into a set $\vec{d}_1$ and a set $\vec{d}_2$. $\vec{d}_1$ and $\vec{d}_2$ may stem for example from measurements on different days. The model which we employ to explain the data may be characterized by a set of parameters $\vec{\theta}$. The question then arises wether we should analyze the composite data set $\vec{d}$ or should we analyze $\vec{d}_1$ and use the posterior distribution of the parameters $\vec{\theta}$ as a prior distribution for the analysis of $\vec{d}_2$. The posterior distribution of the parameters given the composite data set is given by Bayes' theorem as

$$p(\vec{\theta}|\vec{d}_1,\vec{d}_2,I) = p(\vec{\theta}|I)\cdot p(\vec{d}_1,\vec{d}_2|\vec{\theta},I)/p(\vec{d}_1,\vec{d}_2|I) \tag{76}$$

Next we expand the likelihood $p(\vec{d}_1,\vec{d}_2|\vec{\theta},I)$ employing the product rule

$$p(\vec{d}_1,\vec{d}_2|\vec{\theta},I) = p(\vec{d}_1|\vec{\theta},I)\cdot p(\vec{d}_2|\vec{d}_1,\vec{\theta},I) \tag{77}$$

If we now assume that the values $\vec{d}_2$ are not influenced by the previous measurements $\vec{d}_1$, then the second factor in (77) becomes logically independent of $\vec{d}_1$. Equation (76) may

therefore be rewritten in the form

$$p(\vec{\theta}|\vec{d}_1,\vec{d}_2,I) = \frac{p(\vec{\theta}|I)p(\vec{d}_1|\vec{\theta},I)}{p(\vec{d}_1|I)} \cdot \frac{p(\vec{d}_2|\vec{\theta},I)}{p(\vec{d}_2|\vec{d}_1,I)} \tag{78}$$

The first factor in (78) is obviously the posterior distribution of the parameters based on the data set $\vec{d}_1$. The second factor is the likelihood of the data set $\vec{d}_2$ divided by the conditional evidence $p(\vec{d}_2|\vec{d}_1,I)$. This term cannot be simplified further since the evidence of the data $\vec{d}_2$ depends of course on the information about the parameters $\vec{\theta}$ contained in data set $\vec{d}_1$. The situation is similar to the previous discussion of the prior predictive value. $p(\vec{d}_2|\vec{d}_1,I)$ will reach its maximum when data set $\vec{d}_2$ is best explained with the numerical values of the parameters from the analysis of data set $\vec{d}_1$. The answer to our questions is therefore that sequential and one step parameter estimation are equivalent and lead to the same answer. Note that our proof assumes that the full posterior distribution based on $\vec{d}_1$ is employed as a prior for the second step. The equivalence of sequential and one step analysis is not guaranteed if only estimates of mean and (co)variance of the posterior from the first step enter as the prior information in the second step of the analysis.

II. ASSIGNING PROBABILITY DISTRIBUTIONS

We have employed Gaussian functions for the prior and the likelihood in chapter I. In the latter case we made the strong assumption of a Gaussian error distribution. Gaussian functions are convenient for analytical calculations. However what do they really mean? Are there other possibilities? We shall now turn to the general problem of formulating distribution functions which encode our knowledge. Two different paths will be pursued: the principle of maximum entropy and the requirement of transformation invariance.

A. The principle of maximum entropy

The derivation of distribution functions from the principle of maximum entropy subject to additional restrictive conditions was first introduced by E.T. Jaynes (Phys. Rev. 106 (1957) 620). The Shannon entropy (C.E. Shannon, Bell Syst. Techn. Journal 27(1948) 379) reads

$$S = -\sum_i p_i \log \frac{p_i}{m_i}, \quad S = -\int p(x) \log \frac{p(x)}{m(x)} dx \tag{79}$$

where p is the distribution whose entropy is to be maximized and m some reference distribution which makes the argument of the log function invariant under a transform of variables. $m(x)$ allows also to introduce prior knowledge which we deem important about $p(x)$. This can be most easily seen if the additional knowledge which we would like to encode about $p(x)$ or rather p_i is the normalization. We treat the discrete case for illustration. We use the formalism of Lagrange multipliers to maximize the entropy subject to the condition $\sum_i p_i = 1$.

$$\begin{aligned} Q &= -\sum_i p_i \log(p_i/m_i) + \lambda \left(\sum_i p_i - 1 \right) \\ \frac{\partial Q}{\partial p_k} &= -\log(p_k/m_k) - 1 + \lambda = 0 \\ p_k &= m_k \cdot \exp\{1-\lambda\}, \quad \sum_i p_i = 1 \end{aligned} \tag{80}$$

The normalization condition fixes the Lagrange parameter λ to be 1. Hence we obtain $p_i = m_i$ for all i.

Let us pursue the discrete case further and consider a dice. In addition to the normalization we now impose the auxiliary condition that the average score in a sufficiently large number of throws is $3.5(1+\varepsilon)$. We will see that $\varepsilon = 0$ corresponds to an ideal dice. However a dice is never really ideally balanced. The question is now what can we say about the probability of a particular outcome of a throw given only the information that the probability is normalized and the average result is $3.5(1+\varepsilon)$? The Lagrange optimization problem is now

$$Q = -\sum_j p_j \log 6 p_j + \mu \left(\sum_j p_j - 1 \right) + \lambda \left(\sum_j j p_j - 3.5(1+\varepsilon) \right) \tag{81}$$

In this functional we have taken $m_j = 1/6$ for all j, a result which we expect for an ideal dice. Proceeding in the optimization as before we obtain the system of equations

$$\begin{aligned} p_k &= \frac{1}{6} \exp(\mu - 1 + \lambda) \\ \sum_k p_k &= 1 \\ \sum_k k \cdot p_k &= 3.5(1+\varepsilon) \end{aligned} \tag{82}$$

Inserting the explicit expression for p_k in the two conditions yields

$$\begin{aligned} \frac{1}{6} \cdot e^{\mu-1} \sum_{k=1}^{6} e^{k\cdot\lambda} &= \frac{1}{6} e^{\mu-1} e^{\lambda} \sum_{k=1}^{6} \left(e^{\lambda}\right)^{k-1} \\ \frac{1}{6} \cdot e^{\mu-1} \sum_{k=1}^{6} k \cdot e^{k\cdot\lambda} &= \frac{1}{6} e^{\mu-1} \frac{\partial}{\partial\lambda} \sum_{k=1}^{6} e^{k\cdot\lambda} \end{aligned} \tag{83}$$

The series in the first equation of (83) is a geometric series with initial term 1 and increment e^{λ} which sums to

$$\sum_{k=1}^{6} p_k = \frac{1}{6} e^{\mu-1} \cdot \frac{e^{7\lambda} - e^{\lambda}}{e^{\lambda} - 1} \tag{84}$$

The calculation of the second equation (83) is now straightforward and yields

$$\sum_{k=1}^{6} k \cdot p_k = \frac{1}{6} e^{\mu-1} \cdot \frac{6e^{8\lambda} - 7e^{7\lambda} + e^{\lambda}}{(e^{\lambda} - 1)^2} \tag{85}$$

The ratio of (85) and (84) eliminates the Lagrange parameter μ and leads to the equation which determines λ

$$3.5(1 + \varepsilon) = \frac{6e^{8\lambda} - 7e^{7\lambda} + e^{\lambda}}{(e^{\lambda} - 1)(e^{7}\lambda - e^{\lambda})} \tag{86}$$

An approximate solution of (83) can be obtained for small ε by expanding numerator and denominator in (86) up to third (!) order with the result

$$\lambda \approx 6 \cdot \varepsilon / 5 \tag{87}$$

The following table presents the solution p_j for an ideal and two slightly misbalanced dices. Tiny violations of the normalization occur in course of rounding the results to three significant figures. The approximation error involved in (87) is smaller in all cases. What is the significance of our results? Surely there exists an infinite manifold of differently unbalanced dices having a preset average score.

av. Score	p_1	p_2	p_3	p_4	p_5	p_6
3.71	0.138	0.149	0.160	0.172	0.184	0.198
3.50	0.167	0.167	0.167	0.167	0.167	0.167
3.29	0.198	0.184	0.172	0.160	0.149	0.138

The principle of maximum entropy selects from this manifold the least informative possibility. This argument applies also to a dice with average score 3.5. Also for this average

score there exits an infinite manifold of possible realizations. Maximum entropy selects from this manifold the uniform distribution $p_j = 1/6$ for all j. So far for maximum entropy in action for the case of discrete probabilities! Next we consider maximizing the entropy of a continuous probability distribution subject to side conditions.

As a first type of condition we consider the knowledge of moments M_k of the function $p(x)$

$$M_k = \int x^k p(x) dx \tag{88}$$

The optimization problem is then

$$Q = -\int p(x) \log \frac{p(x)}{m(x)} + \sum_k \lambda_k \left\{ \int x^k p(x) dx - M_k \right\} \tag{89}$$

The functional derivative of Q with respect to $p, \delta Q$ is called the variation

$$\delta Q = -\int \delta p \left\{ \log \frac{p(x)}{m(x)} + 1 - \sum_k \lambda_k x^k \right\} dx = 0 \tag{90}$$

Since we require that this integral (90) vanishes for arbitrary variations δp the conclusion is that the bracket under the integral in (90) must vanish. This yields immediately

$$p(x) = m(x) \cdot \exp \left\{ -1 + \sum_k \lambda_k x^k \right\} \tag{91}$$

and λ_k have to be chosen such that

$$M_k = \int m(x) x^k \exp \left\{ -1 + \sum_i \lambda_i x^i \right\} \tag{92}$$

Note that M_k can be obtained from M_0 via

$$M_k = \frac{\partial}{\partial \lambda_k} M_0 \tag{93}$$

We shall now consider important special cases. Suppose our only knowledge consist of the normalization requirement in a support interval $[a, b]$. Assume further that $m(x) = 1$. Then

$$M_0 = \int_a^b \exp \{-1 + \lambda_0\} \, dx = (b - a) \exp \{-1 + \lambda_0\} = 1 \tag{94}$$

$$p(x) = \exp \{-1 + \lambda_0\} = \frac{1}{b - a} \tag{95}$$

The principle of maximum entropy yields in this case the uniform distribution. The result is intuitively correct and in line with Bernoulli's "principle of insufficient reason" later on renamed by Kaynes as the "principle of indifference". A second important special case arises if we assume to know in addition to normalization also the expectation value $< x >$ of x in $[a, b]$.

$$M_0 = \int_a^b \exp\{-1 + \lambda_0 + \lambda_1 x\}\, dx$$

$$M_0 = \exp\{-1 + \lambda_0\}\, \frac{1}{\lambda_1} \cdot (\exp(\lambda_1 b) - \exp(\lambda_1 a)) \tag{96}$$

The first moment is obtained using (93) as

$$M_1 = \frac{e^{-1+\lambda_0}}{\lambda_1} \left\{ b e^{\lambda_1 b} - a e^{\lambda_1 a} - (e^{\lambda_1 b} - e^{\lambda_1 a})/\lambda_1 \right\} \tag{97}$$

No general analytical solution of (96, 97) in terms of λ_0, λ_1 exists. For arbitrary numbers a, b the equations must be solved numerically. This affects the numerical values of λ_0, λ_1 but not the fact that the distribution is an exponential. The important special case $a = 0$ and $b \to \infty$ has however a simple analytical solution and leads to

$$p(x| < x >) = \frac{1}{< x >} \cdot \exp\{-x/ < x >\} \tag{98}$$

Let us finally consider the case that M_0, M_1, and M_2 are known numbers. The maximum entropy distribution is in this case

$$p(x) = m(x) \cdot \exp\left\{-1 + \lambda_0 + \lambda_1 x + \lambda_2 x^2\right\} \tag{99}$$

a shifted Gaussian regardless of what the support interval of x is . For the special case $m \equiv 1$ and $-\infty < x < \infty$ the integral M_0 has already been calculated previously (28)

$$M_0 = \left(\frac{\pi}{\lambda_2}\right)^{1/2} \exp\left\{+1 - \lambda_0 - \lambda_1^2/4\lambda_2\right\} \tag{100}$$

Expressions for M_1 and M_2 are then generated by application of (93). In terms of the characteristic parameters of a Gaussian, namely position $< x >= M_1/M_0$ and variance $\sigma^2 = (M_2/M_0 - (M_1/M_0)^2)$ we obtain

$$p(x| < x >, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x - < x >)^2\right\} \tag{101}$$

This result (101) is very important, because it tells us that only knowledge of the first three moments and application of the principle of maximum entropy specify the least informative probability distribution as a Gaussian. Our strong assumption in formulating the likelihood (44) that the errors are Gaussian distributed can therefore be relaxed to the assumption that the mean error ε is equal to $<\varepsilon_i>=0$ and the mean square error is $<\varepsilon_i^2>=\sigma^2$.

Among the infinite number of distributions which can be chosen to fulfil these criteria, for example rectangular or triangular distributions on a finite support, the principle of maximum entropy selects a Gaussian. The Gaussian likelihood (44) applies therefore to a much wider class of problems as initially expected. Having exploited the principle of maximum entropy for the ordinary moments M_k (88) we now consider the generalized moment

$$M^g(p) = \int g(x)p(x)dx \tag{102}$$

We maximize

$$\begin{aligned} \phi \; &= -\int p(x)\log\frac{p(x)}{m(x)}dx + \lambda\left(M^g - \int g(x)p(x)dx\right) \\ \delta\phi = & \qquad \int \delta p\left\{-\log\frac{p(x)}{m(x)} - 1 - \lambda g(x)\right\}dx \end{aligned} \tag{103}$$

For arbitrary nontrivial variations of p the bracket under the integral must vanish. This yields

$$\begin{aligned} p(x) &= m(x)\cdot\exp\{-1-\lambda g(x)\} \\ M^g &= \int m(x)\exp\{-1-\lambda g(x)\}\, g(x)dx \end{aligned} \tag{104}$$

Important functionals $g(x)$ are

$g_1(x) = |f(x)|^2, \quad g_2(x) = |f'(x)|^2, \quad g_3(x) = |f''(x)|^2, \quad g_4(x) = f'^2/f$

which quantify our knowledge about the "power" contained in $f(x), (g_1)$, or characterize the smoothness of $f(x)$ in different ways (g_2, g_3, g_4).

B. Prior probabilities by transformation invariance.

The principle of maximum entropy provides a powerful recipe to formulate prior probabilities which encode specific knowledge. Another important class of priors, in general less informative, results from the requirement of transformation invariance. We shall consider in the following specific cases. Let σ be a scale variable as for example in (44) where σ sets the

scale for $\{\varepsilon_i\}$. Let $p(\sigma)$ be the probability distribution for σ. $P(\sigma)$ is a probability density and $p(\sigma)d\sigma$ is an infinitesimal mass element of probability and is a number. Transformation invariance requires that this mass element remains the same when described in another system of coordinates σ'. For the number $p(\sigma)d\sigma$ this implies

$$p(\sigma)d\sigma = p(\sigma')d\sigma' \tag{105}$$

but of course $p(\sigma) \neq p(\sigma')$ in general. Let us now investigate the behaviour of (105) under the transform $\sigma' = \alpha \cdot \sigma, \quad d\sigma = \alpha d\sigma$

$$p(\sigma)d\sigma = p(\alpha\sigma) \cdot \alpha d\sigma \tag{106}$$

(106) must also hold for infinitesimal transformations. To exploit this requirement we Taylor expand both sides of (106) around $\alpha = 1$. This yields

$$p(\sigma) = p(\sigma) + \{p(\alpha\sigma) + \alpha \cdot \sigma p'(\alpha\sigma)\}_{\alpha=1} (\alpha - 1) \tag{107}$$

Comparing coefficient of powers of $(\alpha - 1)$ on both sides leads to

$$p(\sigma) + \sigma p'(\sigma) = 0 \tag{108}$$

with the solution

$$p(\sigma) = \frac{const}{\sigma} \tag{109}$$

The distribution (109) is known as Jeffrey's prior. It is not normalizable in $0 \leq \sigma \leq \infty$. A distribution which is not normalizable is called improper. Of course there does not exist any moment. The distribution is therefore also uninformative. Improper distributions require special care, they should always be regarded as the limit of a sequence of proper distributions. (109) is normalizable in any finite interval $1/B \leq \sigma \leq B$ and is therefore the limit of the sequence

$$p(\sigma) = \frac{1}{2\log B} \cdot \frac{1}{\sigma}, \quad \frac{1}{B} \leq \sigma \leq B \tag{110}$$

for $B \to \infty$. Before proceeding to another special case we shall demonstrate an application of (110).

C. Application of Jeffrey's prior.

When formulating the likelihood function for a data set $\{d_i\}$ with common mean μ we have assumed in (44) that all the data have the same true error σ. While we may have

reason to assume that the measurement error on all the data is the same we hardly ever know its true value. We now take the opposite also unrealistic position to pretend to be entirely ignorant about the true error σ. In this case we must infer the mean value from the marginal likelihood $p(\vec{d}|\mu, I)$ instead from $p(\vec{d}|\mu, \sigma, I)$. $p(\vec{d}|\mu, I)$ is obtained by application of the marginalization rule as

$$p(\vec{d}|\mu, I) = \int p(\vec{d}, \sigma|\mu, I)d\sigma = \int p(\sigma|\mu, I)p(\vec{d}|\mu, \sigma, I)d\sigma \tag{111}$$

Of course $p(\sigma|\mu, I)$ is logically independent of μ. Complete ignorance with respect to σ is expressed by (110)

$$p(\vec{d}|\mu, I) = \left(\frac{1}{2\pi}\right)^{N/2} \frac{1}{Z_\varepsilon} \int_{1/B}^{B} \left(\frac{1}{\sigma}\right)^{N-\varepsilon} \exp(-\frac{\alpha}{\sigma^2})\frac{d\sigma}{\sigma} \tag{112}$$

where we have introduced a new parameter ε. It serves to incorporate Jeffrey's prior for $\varepsilon = 0$ or alternatively a constant uniform prior in $1/B \leq \sigma \leq B$ for $\varepsilon = 1$. The prior normalization depends of course on ε and this is expressed by the $\varepsilon-$index on Z. The exponent α is an abbreviation for

$$\alpha = \frac{1}{2}\sum_i (d_i - \mu)^2 = \frac{1}{2}\sum_i (d_i - \overline{d} + \overline{d} - \mu)^2 =$$
$$\frac{1}{2}\sum_i (d_i - \overline{d})^2 + \frac{1}{2}\sum_i (\overline{d} - \mu)^2 + \sum_i (d_i - \overline{d})(\overline{d} - \mu) \tag{113}$$

The second sum on the right hand side does not depend on i and is hence equal to $1/2 \cdot N(\overline{d} - \mu)^2$. In the third term we can take for the same reason $(\overline{d} - \mu)$ out of the sum which is then seen to be equal to zero. Then α becomes

$$\alpha = \frac{1}{2}N(\overline{d} - \mu)^2 + \frac{1}{2}\sum_i (d_i - \overline{d})^2 \tag{114}$$

The above integral (112) shows up frequently in Bayesian calculations. It can be obtained in closed form for $B \to \infty$ by substituting $x = \alpha/\sigma^2$

$$I_N(\alpha) = \int_0^\infty \left(\frac{1}{\sigma}\right)^N \exp(-\alpha/\sigma^2)\frac{d\sigma}{\sigma} = \frac{1}{2} \cdot \frac{\Gamma(\frac{N}{2})}{\alpha^{N/2}} \tag{115}$$

Collecting terms we arrive at the marginal likelihood

$$p(\vec{d}|\mu, I) = \left(\frac{1}{\pi}\right)^{N/2} \cdot \frac{1}{2} \quad \frac{\Gamma\left(\frac{N-\varepsilon}{2}\right)}{\left\{N(\overline{d} - \mu)^2 + \sum_i (d_i - \overline{d})^2\right\}^{\frac{N-\varepsilon}{2}}} \tag{116}$$

Our goal, an estimate of μ and its confidence limit follows from Bayes' theorem

$$p(\mu|\vec{d}, I) = p(\mu|I) \cdot p(\vec{d}|\mu, I)/p(\vec{d}|I) \tag{117}$$

Choosing $p(\mu|I)$ flat in some sensible range $\mu_{min} \leq \mu \leq \mu_{max}$ completes the specification of the posterior distribution. Note, that it is non Gaussian and consequently our considerations about characterization of the posterior apply. This yields the point estimate $\hat{\mu}$

$$\frac{d}{d\mu} \log p(\mu)|_{\mu=\hat{\mu}} = 0, \qquad \hat{\mu} = \overline{d} \tag{118}$$

and the variance

$$\frac{1}{\sigma^2} = -\frac{d^2}{d\mu^2} \log p(\mu)|_{\hat{\mu}} = \frac{N(N-\varepsilon)}{\sum\limits_i (d_i - \overline{d})^2} \tag{119}$$

Thus the posterior is summarized by

$$\hat{\mu} = \overline{d} \pm \sqrt{\frac{\sum (d_i - \overline{d})^2}{(N-\varepsilon)N}} \tag{120}$$

The to most physicists familiar $N(N-1)$ denominator under the square root results if we assume a flat prior in σ. The appropriate uninformative prior is however $1/\sigma$ corresponding to $\varepsilon = 0$. We realize that this results in a slightly smaller variance and represents correctly the inference on the basis of the assumed prior knowledge.

D. Location prior

If we apply to (105) instead of the similarity transformation $\sigma' = \alpha\sigma$ the translation $\mu' = \mu + b$ then

$$p(\mu)d\mu = p(\mu + b)d\mu' \tag{121}$$

$d\mu = d\mu'$ in this case. We Taylor expand both sides of (121) as before around $b = 0$

$$p(\mu) = p(\mu) + p'(\mu)(\mu - b) \tag{122}$$

and obtain after comparison of coefficients

$$p'(\mu) = 0 \qquad\qquad p(\mu) = const \tag{123}$$

Again, this prior for a location variable is only normalizable on a finite support. If we insist on complete ignorance the support is infinite and the prior is improper. For practical

purposes it should always be regarded as the limit of

$$p(\mu|B) = \frac{1}{2B}, \qquad -B \leq \mu \leq B, \qquad B \to \infty \tag{124}$$

or other conceivable sequences. From the foregoing two examples one might wonder wether transformation invariance leads invariably to improper priors . This is not the case as we shall demonstrate now.

E. Straight line fit

An interesting and in the analysis of physics data ubiquitous problem is the fit of a straight line to a collection of points (x_i, y_i). We shall first deal with the simplest possible model which consists of straight lines passing through the origin. Assume the model

$$y_i = ax_i + \varepsilon_i \tag{125}$$

and in addition $< \varepsilon_i >= 0$ and $< \varepsilon_i^2 >= \sigma^2$. The likelihood for this case follows then from the principle of maximum entropy as

$$p(\vec{y}|\vec{x}, \sigma, a, I) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - ax_i)^2\right\} \tag{126}$$

from which we want to estimate the slope a. Bayes' theorem allows for this inverse conclusion

$$p(a|\vec{y}, \vec{x}, \sigma, I) = p(a|I)p(\vec{y}|\vec{x}, \sigma, a, I)/p(\vec{y}|\vec{x}, \sigma, I) \tag{127}$$

In order to exploit (127) we need to assign a prior in a. This is a nontrivial problem as is evident from Fig. 5. The left panel shows a collection of straight lines through the origin with constant increment in slope. This corresponds to a flat prior in a. This is obviously unacceptable since this prior contains the information of an amassment versus large slopes. Intuitively an uninformative prior for a straight line through the origin should attach equal probability to every direction such as in the right panel of Fig. 5. This figure was generated with constant increments in angle with respect to the x-axis. In other words it is flat in the polar angle φ. The important lesson to be learned from this figure is that **a distribution which is flat in one variable need not remain flat after a transformation of that variable**, in our case $a = tg\varphi$. The flat prior in φ is, properly normalized

$$p(\varphi) = \frac{1}{\pi} \qquad , \quad -\frac{\pi}{2} \leq \varphi \leq \frac{\pi}{2} \tag{128}$$

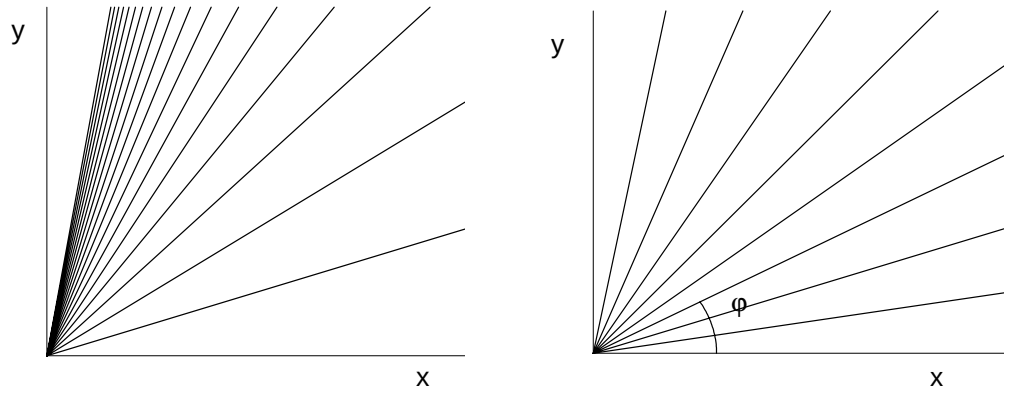

FIG. 5: The left hand panel shows straight lines through the origin with constant increment Δa in slope. The right hand panel shows a series of straight lines with constant increment $\Delta\varphi$ of the angle versus the x-axis.

In course of a transformation $a = tg\varphi$ we have

$$\begin{aligned} & p(\varphi)d\varphi = p(a)da \\ & p(a) = p(\varphi) \cdot \left|\frac{d\varphi}{da}\right| = \frac{1}{\pi}\left|\frac{d\varphi}{da}\right| \\ & a = tg\varphi \quad , \quad \frac{da}{d\varphi} = \frac{1}{\cos^2\varphi} \quad , \quad \cos\varphi = (1+a^2)^{-1/2} \end{aligned} \tag{129}$$

$$p(a) = \frac{1}{\pi}\cdot\frac{1}{1+a^2} \tag{130}$$

This result, which is based on our intuition of an uninformative prior for the orientation of a straight line through the origin, shall now be derived from the transformation invariance requirement. We consider the two systems of coordinates (x, y) and (x', y') which are rotated by an angle φ with respect to each other (Fig.6).

The angle of the straight line with the x-axis is α and with the x' axis β. Then

$$y = ax \quad , \quad a = tg\alpha \qquad , \qquad y' = a'x' \quad , a' = tg\beta \tag{131}$$

α, β and φ are related through the addition theorem for the tg function

$$tg\beta = \frac{tg\alpha - tg\varphi}{1 + tg\alpha\, tg\varphi} \tag{132}$$

If we define $z = tg\varphi$ and $a' = (a - z)/(1 + az)$ then

$$p(a)da = p(a')da' = p\left(\frac{a-z}{1+az}\right)da' \tag{133}$$

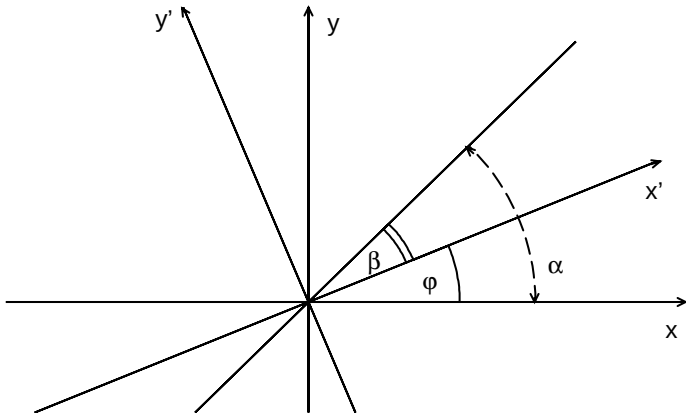

FIG. 6:

since $p(a)da$ and $p(a')da'$ are the same elements of probability mass described in two different systems of coordinates. The Jacobian is

$$\left|\frac{da}{da'}\right| = \frac{(1+az)^2}{1+z^2} \tag{134}$$

and hence we obtain for all z

$$p(a)\frac{(1+az)^2}{1+z^2} \quad = \quad p\left(\frac{a-z}{1+az}\right) \tag{135}$$

Taylor expansion of both sides of (132) around $z = 0$ yields

$$p(a) + p(a)\frac{d}{dz}\left.\frac{(1+az)^2}{1+z^2}\right|_{z=0} \cdot z = p(a) + p'(a)\frac{d}{dz}\left.\frac{a-z}{1+az}\right|_{z=0} \cdot z \tag{136}$$

Comparing the coefficients of z yields the differential equation

$$2ap(a) = -(1+a^2)p'(a) \tag{137}$$

with properly normalized solution

$$p(a) = \frac{1}{\pi} \quad \frac{1}{1+a^2} \tag{138}$$

which is the same as our intuitive guess (130).

The obvious and practically important generalization of our model (125) introduces a constant offset b

$$y_i = \alpha x_i + b + \varepsilon_i \tag{139}$$

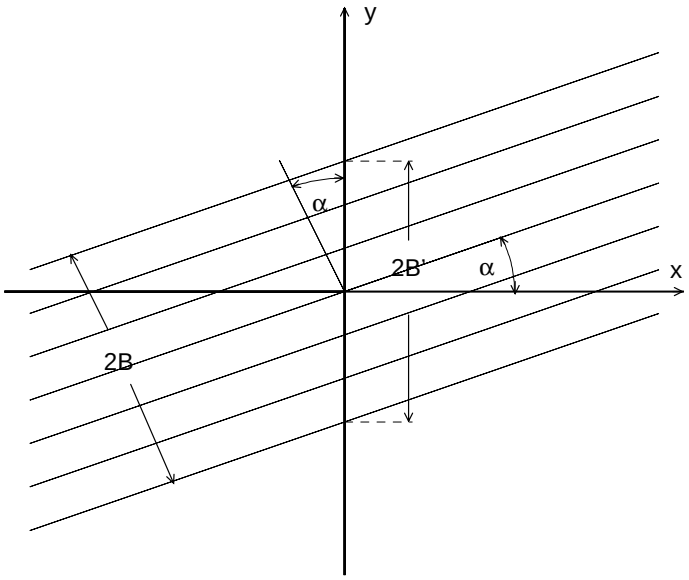

FIG. 7:

The likelihood for this model is

$$p(\vec{y}|\vec{x},\sigma,a,b,I) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - ax_i - b)^2\right\} \tag{140}$$

from which we wish to estimate in addition to slope a also the offset b. Bayes' theorem yields

$$p(a,b|\vec{y},\vec{x},\sigma,I) = p(a,b|I)\cdot p(\vec{y}|\vec{x},\sigma,a,b,I)/p(\vec{y}|\vec{x},\sigma,I) \tag{141}$$

$p(a,b|I)$ can be further expanded into $p(a|I)\cdot p(b|a,I)$ with $p(a|I)$ already known (138). To obtain $p(b|a,I)$ consider a bundle of parallel straight lines of width 2 B. The associated density is (see Fig.7)

$$p(b|a=0,I) = 1/2B \tag{142}$$

For arbitrary a the width of the bundle is $2B'$ with density $1/2B'$, hence

$$p(b|a,I) = \frac{1}{2B'} = \frac{\cos\alpha}{2B} = \frac{1}{2B}\cdot\frac{1}{\sqrt{1+a^2}} \tag{143}$$

The complete, properly normalized prior for a,b in $-\infty < a < \infty$ and $-B \le b \le B$ is then

$$p(a,b|I) = \frac{1}{4B}\cdot\frac{1}{(1+a^2)^{3/2}} \tag{144}$$

This completes our examples for the application of the requirement of transformation invariance to the determination of prior probabilities.

III. PARAMETER ESTIMATION

Data from physics experiments are usually explained by a model which in turn is specified by a certain set of parameters. The particular numerical values of the parameters select one particular model out of the whole class. The traditional way to estimate these numerical values on the basis of the given data has been the least squares method. Bayesian estimation goes beyond the least squares approach because it allows to incorporate prior knowledge into the analysis. The important advantage of Bayesian parameter estimation using informative priors is that any new bit of information can be processed. In particular progress can be made also if the number of data is smaller than the number of parameters to be estimated. In this section we shall apply the results of Chapter I and II in two specific examples of parameter estimation.

A. The weighted arithmetic mean

The first example is again the arithmetic mean of a series of data d_i with (unknown) true errors σ_i and apparent - as measured - errors s_i. We shall be interested in the estimate of the mean and shall further address the question wether the quoted errors s_i are consistent with each other. Let μ be the quantity we want to estimate, then

$$p(\vec{d}|\mu,\vec{\sigma},I) = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\sum_i^N (d_i-\mu)^2/\sigma_i^2\right\} \tag{145}$$

Unlike in previous considerations of the arithmetic mean we have assumed in (145) that the true error σ_k may be different for every data point d_k. The likelihood (145) is formulated in terms of the true errors σ_i. These are hardly ever known. Frequently, however, an approximation s_i to the true standard deviation σ_i is known. We shall assume in the following that s_i are known, but may suffer from a systematic deviation from σ_i which we specify by a common factor α. The probability density of σ_i given s_i and α is then

$$p(\sigma_i|s_i,\alpha,I) = \delta(\sigma_i - \alpha s_i) \tag{146}$$

As a first step of our analysis we remove σ_i from the likelihood (145) with the help of (146)

$$p(\vec{d}|\mu,\alpha,\vec{s},I) = \int p(\vec{\sigma}|\vec{s},\alpha,I)\cdot p(\vec{d}|\mu,\vec{\sigma},I)d\vec{\sigma} \tag{147}$$

Integration over σ_i results here simply in the replacement of σ_i by αs_i in (145)

$$p(\vec{d}|\mu,\alpha,\vec{s},I) = \frac{1}{\alpha^N}\left(\prod_i \frac{1}{s_i\sqrt{2\pi}}\right)\exp\left\{-\frac{1}{2\alpha^2}\sum_i (d_i-\mu)^2/s_i^2\right\} \tag{148}$$

We rewrite the exponent in (148) as

$$\sum_i^N (d_i-\mu)^2/s_i^2 = \sum_i^N d_i^2/s_i^2 - 2\mu\sum_i^N d_i/s_i^2 + \mu\sum_i^N 1/s_i^2 \tag{149}$$

It is convenient to introduce the definitions

$$\sum_i^N 1/s_i^2 = N/\rho^2, \sum_i^N d_i/s_i^2 = N\overline{D}/\rho^2, \sum_i^N d_i^2/s_i^2 = N\overline{D^2}/\rho^2 \tag{150}$$

and find the likelihood $p(\vec{d}|\mu,\alpha,\vec{s},I)$

$$p(\vec{d}|\mu,\alpha,\vec{s},I) = \frac{1}{\alpha^N}\left(\prod_i \frac{1}{s_i\sqrt{2\pi}}\right)\exp\left\{-\frac{N}{2\alpha^2\rho^2}(\mu^2 - 2\mu\overline{D} + \overline{D^2})\right\} \tag{151}$$

In the first step we wish to estimate α from this likelihood. Bayes' theorem

$$p(\alpha|\vec{d},\vec{s},I) = p(\alpha|I)\cdot p(\vec{d}|\alpha,\vec{s},I)/p(\vec{d}|\vec{s},I) \tag{152}$$

requires the marginal likelihood $p(\vec{d}|\alpha,\vec{s},I)$ which we obtain from (151) by marginalizing over μ with some suitable prior. Since μ is a location parameter we choose the prior in μ flat as

$$p(\mu|I) = \frac{1}{W}, \quad \mu_{min} \le \mu \le \mu_{max}, \qquad W = \mu_{max} - \mu_{min} \tag{153}$$

The required marginal likelihood is then

$$p(\vec{d}|\alpha,\vec{s},I) = \int_{\mu_{min}}^{\mu_{max}} p(\mu|I)\cdot p(\vec{d}|\mu,\alpha,\vec{s},I)d\mu \tag{154}$$

The likelihood under the integral (154) is more or less strongly peaked as a function of μ. The amount of localization depends on the number and quality of the data. We use the assumption that the integrand in (154) is well localized to replace the limits of integration (μ_{min},μ_{max}) by $\pm\infty$ with negligible effect on the value of the integral. The latter is then again of the standard Gaussian type (22,28).

$$p(\vec{d}|\alpha,\vec{s},I) = \frac{1}{W}\cdot\left(\prod_i \frac{1}{s_i\sqrt{2\pi}}\right)\frac{\sqrt{2\pi}}{\alpha^N}\left\{\frac{\alpha^2\rho^2}{N}\right\}^{1/2}\exp\left\{-\frac{N}{2\alpha^2\rho^2}\overline{\Delta D^2}\right\} \tag{155}$$

Since α^2 sets the scale for $\overline{\Delta D^2}$ in (155) the appropriate uninformative prior for substitution into Bayes' theorem is Jeffrey's prior $1/\alpha$. This yields the evidence

$$p(\vec{d}|\vec{s},I) = \frac{\rho}{W}\cdot\sqrt{\frac{2\pi}{N}}\cdot\left(\prod_i \frac{1}{s_i\sqrt{2\pi}}\right)\int_0^\infty \frac{d\alpha}{\alpha}\frac{1}{\alpha^{N-1}}\exp\left\{-\frac{\phi^2}{\alpha^2}\right\} \tag{156}$$

The remaining integral is of our second standard type (115) and the evidence becomes finally

$$p(\vec{d}|\vec{s},I) = \frac{\rho}{W}\cdot\sqrt{\frac{2\pi}{N}}\left(\prod_i \frac{1}{s_i\sqrt{2\pi}}\right)\cdot\frac{1}{2}\cdot\frac{\Gamma\left(\frac{N-1}{2}\right)}{\phi^{N-1}} \tag{157}$$

This is all we need to write down the posterior distribution of α (152) from which we want to derive estimates for α and $\Delta\alpha^2$. Note that the posterior as a function of α is not a Gaussian. So we expect differences between the exact moments of α, $<\alpha^n>$ which can easily be calculated in the present case and the corresponding moments in Gaussian approximation. The latter are given by (69, 70) as

$$\frac{d}{d\alpha}\log p(\alpha|\vec{d},\vec{s},I) = \left(\frac{2}{\alpha^3}\phi^2 - \frac{N}{\alpha}\right)|_{\alpha=\alpha_{max}} = 0 \tag{158}$$

with solution $\alpha_{max} = \phi\cdot\sqrt{2/N}$

$$\frac{1}{\Delta\alpha_G^2} = -\left(-\frac{6\phi^2}{\alpha^4}+\frac{N}{\alpha^2}\right)|_{\alpha=\alpha_{max}} \quad = N^2/\phi^2 \tag{159}$$

It is interesting to compare the approximate moments with the exact moments in the present case

$$p(\vec{d}|\vec{s},I) <\alpha^2> = \frac{\rho}{W}\sqrt{\frac{2\pi}{N}}\left(\prod_i \frac{1}{s_i\sqrt{2\pi}}\right)\cdot\frac{1}{2}\quad\frac{\Gamma\left(\frac{N-3}{2}\right)}{\phi^{N-3}} \tag{160}$$

$$p(\vec{d}|\vec{s},I) <\alpha> = \frac{\rho}{W}\sqrt{\frac{2\pi}{N}}\left(\prod_i \frac{1}{s_i\sqrt{2\pi}}\right)\cdot\frac{1}{2}\quad\frac{\Gamma\left(\frac{N-2}{2}\right)}{\phi^{N-2}} \tag{161}$$

Combining the results of (157, 160, 161) we obtain for the mean and variance of α in units of ϕ

$$<\alpha>/\phi = \frac{\Gamma\left(\frac{N-2}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)},\quad <\Delta\alpha^2>/\phi^2 = \left\{\frac{2}{N-3}-\left[\frac{\Gamma\left(\frac{N-2}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}\right]^2\right\} \tag{162}$$

The following table compares this result (162) with the Gaussian approximation (158,159). $<\alpha>$ and $<\Delta\alpha^2>$ have again been taken in units of ϕ and ϕ^2 respectively. These quantities are unique functions of the sample size N. The table shows that the Gaussian approximation can be considerably in error particularly for small samples. It is by no means

	approximate		exact	
N	α_{max}	$\Delta\alpha_G^2$	$<\alpha>$	$<\Delta\alpha^2>$
4	0.70711	0.06250	1.12838	0.72676
8	0.50000	0.01563	0.60180	0.03783
16	0.35355	0.00391	0.38477	0.00580
32	0.25000	0.00098	0.26036	0.00118

a panacea. The table suggests, however, that the difference between exact and approximate values vanishes rapidly as the sample size becomes large. This may even be proven if we replace the Γ-function ratios in (162) by their asymptotic values employing Stirling's formula

$$\Gamma(z) \sim e^{-z} z^{z-1/2} \cdot \sqrt{2\pi} \tag{163}$$

This yields the ratio for large N

$$\Gamma\left(\frac{N-2}{2}\right) / \Gamma\left(\frac{N-1}{2}\right) \sim e^{1/2} \cdot \sqrt{\frac{2}{N-2}} \cdot \left(\frac{N/2-1}{N/2-1/2}\right)^{N/2-1} \sim \sqrt{\frac{2}{N}} \tag{164}$$

Substituting $x/2 = N/2 - 1$ in the rightmost bracketed term this may be shown to be equal to $e^{-1/2}$, and the large N limit of $<\alpha>$ becomes equal to the Gaussian approximation result.

The second part of our parameter estimation exercise is the estimate of μ and its variance $\Delta\mu^2$. The posterior distribution $p(\mu|\vec{d},\vec{s},I)$ is given by Bayes' theorem

$$p(\mu|\vec{d},\vec{s},I) = p(\mu|I)p(\vec{d}|\mu,\vec{s},I)/p(\vec{d}|\vec{s},I) \tag{165}$$

in terms of the marginal likelihood $p(\vec{d}|\mu,\vec{s},I)$ and the prior $p(\mu|I)$, for which we shall again choose (153), and the evidence $p(\vec{d}|\vec{s},I)$ which we have already calculated (157). It remains to calculate the integrals

$$p(\vec{d}|\vec{s},I) \cdot <\mu^k> = \int \mu^k d\mu\, p(\mu,I) \int \frac{d\alpha}{\alpha} p(\vec{d}|\mu,\alpha,\vec{s},I) \tag{166}$$

Since $p(\vec{d}|\mu,\alpha,\vec{s},I)$ is of Gaussian type in μ we conveniently reverse the order of integration and apply again the standard Gaussian formulae. The subsequent α-integration leads then to formulae of the standard type (115). The procedure is somewhat lengthy but elementary and yields

$$<\mu> = \overline{D} \qquad , \qquad <\Delta\mu^2> = \frac{\overline{\Delta D^2}}{N-3} \tag{167}$$

Recall the definition of $\overline{D^2}$ and $\overline{D}$ (129) to realize that the variance $< \Delta\mu^2 >$ depends now on the errors of the individual measurements s_i **and** on the data dispersion.

Our result differs in this respect from the conventional well known weighted average which we obtain from (151) by fixing α to $\alpha = 1$. This corresponds to the state of knowledge that the measured errors s_i are identical to the true errors σ_i (146) in which case

$$< \mu|\alpha = 1 >= \overline{D} \qquad , < \Delta\mu^2|\alpha = 1 >= \frac{\rho^2}{N} = \left\{ \sum 1/s_i^2 \right\}^{-1} \tag{168}$$

It is immediately obvious that our result (167) is much superior to (168) since it is intuitively unacceptable that the standard deviation of the mean should only depend on the quoted errors of the contributing data regardless of the data scatter.

We shall now discuss an application of the developed formulae. The system of physical constants and conversion factors arises from a critical joint evaluation of a series of measurements like the quantum Hall effect (QHE), calibration of the Ampère (A), gyromagentic ratio of the proton γ_p and many more. The employed algorithm is essentially least squares, which means that the errors quoted with the individual measurements determine the weight of the respective quantity in the overall evaluation. Out of the many possibilities let us consider the raw data on the quantum Hall effect and the calibration factor for the realization of the Ampère and answer the question wether the data are consistent with their quoted error.

The following table is drawn from Rev.Mod. Phys. 59 (1121) 1987.

As a result of our analysis to the above data we realize that the conventional error estimate (168) is too optimistic for the quantum hall effect while the quoted error for the calibration factor of the Ampère is unduly large. This means with respect to the evaluation of physical constants and conversion factors that the weight of the quantum hall effect data is too large by a factor of $(21/16)^2 = 1 : 1.7$ while the calibration factor for the Ampère deserves an increased weight by $(25/20)^2 = 1.25$. This concludes our first example for parameter estimation.

QHE		A	
value	error	value	error
25812.8469	0.0048	0.9999974	0.0000084
25812.8495	0.0031	0.9999982	0.0000059
25812.8432	0.0040	0.9999986	0.0000061
25812.8427	0.0034	1.0000027	0.0000097
25812.8397	0.0057	1.0000032	0.0000079
25812.8502	0.0039	1.0000062	0.0000041
mean	error	mean	error
	0.0016 [168]		0.0000025 [168]
25812.8461		1.0000021	
	0.0021 [167]		0.0000021 [167]

B. Straight line fit

As a second example we choose the fit of a linear model $y = ax+b$ to a set of data (x_i, y_i). The model equation is

$$y_i - ax_i - b = \varepsilon_i, \quad i = 1, ..., N \tag{169}$$

We assume that $< \varepsilon_i >= 0$ and $< \varepsilon_i^2 >= \sigma^2$ are known. The maximum entropy distribution based on this knowledge is

$$p(\vec{d}|\vec{x}, a, b, \sigma, I) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\sum_i^N (y_i - ax_i - b)^2\right\} \tag{170}$$

Our goal is to estimate parameters a, b from the likelihood (170) via Bayes' theorem

$$p(a, b|\vec{y}, \vec{x}, \sigma, I) = p(a, b|I) \cdot p(\vec{y}|\vec{x}, a, b, \sigma, I)/p(\vec{y}|\vec{x}, \sigma, I) \tag{171}$$

The prior probability $p(a, b|I)$ was derived previously from transformation invariance requirements (144) so that we can proceed to define the integrals to be done. We treat the evidence first

$$p(\vec{y}|\vec{x}, \sigma, I) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \cdot \frac{1}{4B} \cdot \int_{-\infty}^{\infty} \frac{da}{(1+a^2)^{3/2}} \int_{-B}^{B} db \exp\left\{-\frac{1}{2\sigma^2}\sum_i (y_i - ax_i - b)^2\right\} \tag{172}$$

The inner b-integral has Gaussian structure with respect to b. With the additional assumption that B is sufficiently large in comparison to the width of the exponential the limits of integration $(-B, B)$ can be shifted to $(-\infty, \infty)$ with negligible effect on the value of the integral. These new limits allow once more the application of the formulae for standard Gaussian integrals (22, 28). We rewrite the exponent in (172)

$$\frac{1}{2\sigma^2}\sum_i (y_i - ax_i - b)^2 = \frac{N}{2\sigma^2}\overline{(y-ax)^2} + \frac{N}{2\sigma^2}b^2 - \frac{2bN}{2\sigma^2}\,(\overline{y-ax}) \tag{173}$$

and apply (28) to obtain

$$\begin{aligned} p(\vec{y}|\vec{x},\sigma,I) &= \left[\frac{1}{\sigma\sqrt{2\pi}}\right]^N \cdot \frac{1}{4B} \cdot \frac{\sigma\sqrt{2\pi}}{\sqrt{N}} \\ &\int_{-\infty}^{\infty} \frac{da}{(1+a^2)^{3/2}} \cdot \exp\left\{-\frac{N}{2\sigma^2}\left[\overline{(y-ax)^2} - \overline{(y-ax)}^2\right]\right\} \end{aligned} \tag{174}$$

The remaining integral cannot be done analytically and we resort therefore to the Gaussian approximation (69)

$$\phi(a) = -\frac{3}{2}\log(1+a^2) - \frac{N}{2\sigma^2}\left[\overline{\Delta y^2} + a^2\overline{\Delta x^2} - 2a\overline{\Delta x \Delta y}\right] \tag{175}$$

Let a_0 be the minimum of $\phi(a)$. a_0 is the solution of

$$a\overline{\Delta x^2} - \overline{\Delta x \Delta y} = \frac{-3a}{1+a^2} \quad \frac{\sigma^2}{N} \tag{176}$$

For reasonable data sets meaning large N and moderate σ the right hand side of (173) becomes very small and the lowest order solution in a becomes

$$a_{LS} = \overline{\Delta x \Delta y}/\overline{\Delta x^2} \tag{177}$$

The index LS means "least squares" and signals that a_{LS} is the value of a where the likelihood (170) attains its maximum. We can use (177) to substitute it in the right hand side of (176) for a first order solution for a_0

$$a_0 \approx a_{LS} - \frac{3\cdot\sigma^2}{N\overline{\Delta x^2}} \quad \frac{a_{LS}}{1+a_{LS}^2} \tag{178}$$

The second derivative of $\phi(a)$ at a_0 is

$$\left.\frac{d^2\phi}{da^2}\right|_{a_0} = \quad -\frac{N\overline{\Delta x^2}}{\sigma^2} - 3\frac{1-a_0^2}{(1+a_0^2)^2} = -\frac{1}{\sigma_0^2} \tag{179}$$

and the approximate value of the evidence integral (172) becomes

$$p(\vec{y}|\vec{x},\sigma,I) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^{N-1} \frac{1}{4B} \cdot \sqrt{\frac{2\pi\sigma_0^2}{N}} \exp\{-\phi(a_0)\} \tag{180}$$

The moments of a, or more precisely, mean and variance of a are then calculated from

$$p(\vec{y}|\vec{x},\sigma,I) \cdot < a^n > \quad = \int_{-\infty}^{\infty} \frac{a^n da}{(1+a^2)^{3/2}} \int_{-B}^{B} db \, \exp\{-E(a,b)\} \tag{181}$$

and correspondingly for b

$$p(\vec{y}|\vec{x},\sigma,I) < b^n >= \int_{-\infty}^{\infty} \frac{da}{(1+a^2)^{3/2}} \int_{-B}^{B} b^n db \, \exp\{-E(a,b)\} \tag{182}$$

The inner $b-$integrals are standard in both cases and the remaining $a-$integration proceeds exactly along the lines demonstrated for the evidence. The results are

$$< a > \approx a_0 \qquad\qquad , < \Delta a^2 > \approx \sigma_0^2 \tag{183}$$

$$< b > \approx \overline{y} - a_0 \overline{x} \quad , < \Delta b^2 > \approx \frac{\sigma^2}{N} \tag{184}$$

This completes our analysis in parameter estimation for a linear model in two dimensions. From the definitions of a_0 and σ_0 we find that the terms originating from the prior distribution $p(a,b|I)$ become negligible in the large N small σ limit which is identical to the well known and frequently applied least squares solution.

IV. BAYESIAN MODEL COMPARISON

Comparison of different physics models given a set of data is an ubiquitous problem in physical data analysis. In particular the question how complicated a model should be in order to satisfactorily explain a given data set is of paramount importance. Evidently a model with N parameters will explain any set of N data pointwise. Nothing has been learned in this case. What we want is a model sufficiently refined to explain the mainstream behaviour and sufficiently simple to avoid the fitting of noise. Bayesian probability theory provides a conclusive solution to this problem.

A. Linear versus nonlinear relationship

To begin let us again consider a set of data d_i measured at settings of an independent variable x_i and suppose we have reason to expect a linear relationship between d_i and x_i of the simple form

$$d_i - ax_i = \varepsilon_i \tag{185}$$

Assuming $< \varepsilon_i > = 0$ and $< \varepsilon_i^2 > = \sigma^2$ the maximum entropy distribution for this model is

$$p(\vec{d}|a, \vec{x}, \sigma, M_1) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\sum_i (d_i - ax_i)^2\right\} \tag{186}$$

A hypothetical model M_2 may consist of replacing x_i by $1 - \exp\{x_i\}$, hence

$$p(\vec{d}|a, \vec{x}, \sigma, M_2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left\{-\frac{1}{2\sigma^2}\sum_i \left(d_i - a(e^{x_i} - 1)\right)^2\right\} \tag{187}$$

The question which we then want to answer on the basis of these likelihoods is, what is the probability of model M_k, $k = 1, 2$ given the data $\{d_i\}$, $\{x_i\}$ and σ. Bayes' theorem yields

$$p(M_k|\vec{d}, \vec{x}, \sigma, I) = p(M_k|I) \cdot p(\vec{d}|\vec{x}, \sigma, M_k, I)/p(\vec{d}|\vec{x}, \sigma, I) \tag{188}$$

Since our interest is usually focused on the question of how much model M_1 is to be preferred over model M_2 we take the odds ratio of (188) evaluated for M_1 and M_2 respectively

$$\frac{p(M_1|\vec{d}, \vec{x}, \sigma, I)}{p(M_2|\vec{d}, \vec{x}, \sigma, I)} = \frac{p(M_1|I)}{p(M_2|I)} \cdot \frac{p(\vec{d}|\vec{x}, \sigma, M_1, I)}{p(\vec{d}|\vec{x}, \sigma, M_2, I)} \tag{189}$$

It is composed of two factors. The first factor $p(M_1|I)/p(M_2|I)$ is called the prior odds. This factor allows to incorporate expert knowledge as to how much model M_1 is to be preferred disregarding the newly available data $\{d_i\}, \{x_i\}, \sigma$. The second factor is called the Bayes factor. It modifies the stated expert knowledge by the information contained in the data. The necessity to specify prior knowledge on the models M_1 and M_2 may at first sight cause uneasiness. In physics, however, the ratio $p(M_1|I)/P(M_2|I)$ is in almost all sensible cases equal to one since the unability to give preference to either model is the unique driving force for the design and performance of a new experiment. The evidence associated with the likelihood (187) is

$$p(\vec{d}|\vec{x}, \sigma, M_k, I) = \int p(a|I) \cdot p(\vec{d}|\vec{x}, \sigma, a, M_k, I) da \tag{190}$$

We shall approximate this integral assuming that the likelihood is a well localized function in a such that the variation of the prior $p(a|I)$ in the region where the likelihood differs significantly from zero modifies the value of the integral (190) insignificantly. Thus

$$p(\vec{d}|\vec{x},\sigma,M_k,I) \approx p(a_0|I)\cdot\int da \quad p(\vec{d}|\vec{x},\sigma,a,M_k,I) \tag{191}$$

where a_0 is the value of a where the likelihood has its maximum: a_0 is obtained from (177,186) as $a_0 = \overline{dx}/\overline{x^2}$. The remaining integral is standard Gaussian and we end up with

$$p(\vec{d}|\vec{x},\sigma,M_1,I) \approx p(a_0|I)\cdot\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \sqrt{\frac{2\pi\sigma^2}{N\overline{x^2}}}\exp\left\{-\frac{N\overline{d^2}}{2\sigma^2}\left(1-\frac{(\overline{dx})^2}{\overline{d^2}\,\overline{x^2}}\right)\right\} \tag{192}$$

So far we have left open the question which prior to choose. The natural choice is of course the prior (130) derived from the requirement of transformation invariance. Suppose for a minute we were less educated and chose a flat prior $1/2A$ in the range $-A \le a \le A$. It is clear from (192) that the choice of the prior affects the value of the evidence, in fact for our special case we have

$$\frac{p(\vec{d}|\vec{x},\sigma,M_1,prior1)}{p(\vec{d}|\vec{x},\sigma,M_1,prior2)} = \frac{\pi(1+a_0^2)}{2A} \tag{193}$$

This tells us that the choice of prior is crucial for the numerical value of the evidence. In particular we realize that improper priors $A \to \infty$ will usually be inappropriate in model comparison problems unless they cancel in the odds ratio. This is the case in our simple introductory example. The evidence for model M_2 may be obtained if we define $x^* = e^x - 1$ from (192) simply by replacing x by x^*. The odds ratio becomes then

$$\frac{p(\vec{d}|\vec{x},\sigma,M_1)}{p(\vec{d}|\vec{x},\sigma,M_2)} = \sqrt{\frac{\overline{x^{*2}}}{\overline{x^2}}}\cdot\exp\left\{-\frac{N\overline{d^2}}{2\sigma^2}\left(\frac{(\overline{dx^*})^2}{\overline{d^2}\,\overline{x^{*2}}}-\frac{(\overline{dx})^2}{\overline{d^2}\,\overline{x^2}}\right)\right\} \tag{194}$$

(194) has been evaluated for the two sets of simulated data shown in Fig.8. The data were generated in both cases as $d_i = \exp(x_i) - 1 +$ noise. The variance of the noise was also the same in both sets. The only difference is in the sequence of the random number generator which was used to impose noise on the data. It is customary to quote the natural logarithm of the odds ratio instead of the odds ratio itself. For the left part of Fig.8 the log-odds is $+0.91$ which indicates a slight preference for model (186) while for the right hand data set we find a log-odds of -5.14 indicating a clear preference for model (187). The question arises at this place how the figures on the log-odds should be interpreted. A classification taken from the literature is given in the table.

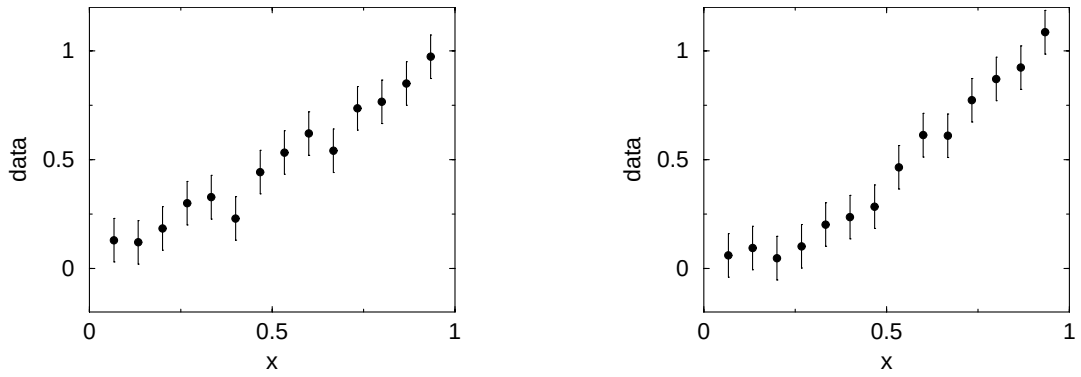

FIG. 8: The data in both panels were generated from $d_i = \exp(x_i) - 1+$ noise. The noise variance is the same in both sets. The only difference between the data in both panels is the random number sequence used to generate the data.

The meaning of log-odds numbers	
hardly significant	< 1
positive evidence	$1 - 2.5$
strong evidence	$2.5 - 5$
overwhelming evidence	> 5

This case of model comparison calls for a comment. Remember, that both data sets in Fig.8 were generated according to $d_i = \exp(x_i) - 1+$ noise. It is the noise imposed on the data which makes up for the difference between the two sets. Probability theory favours a linear model for the left panel though the generator was non-linear. This tells us quite clearly that probability theory can not do wonders. It has no knowledge about the generator and interprets the particular given data sets in terms of the suggested models.

It may not be obvious to the naked eye but can clearly be demonstrated with the help of a ruler that the data in the right hand panel exhibit a slight curvature and are hence preferably explained by model (187).

B. Classification of model comparison

The general problem of Bayesian model comparison can be classified in terms of three cases.

1. The models which we wish to compare contain the same amount of parameters with the same meaning and prior distributions (194, 187, 186).

2. The most general case arises when the number of parameters differs from model to model and different models occupy different volumes in parameter space.

3. An interesting case occurs with hierarchical models. We call a class of models hierarchical if the model of order $(N-1)$ is completely contained in the model of order N. A frequently encountered example for this case is a series representation of an unknown function.

The following approximate consideration can be made in case 3. The evidence is

$$p(\vec{d}|\vec{x},\sigma,M_1,I) = \int p(\vec{d},\vec{\theta}|\vec{x},\sigma,M_1,I)d\vec{\theta} = \int p(\vec{\theta}|I)p(\vec{d}|\vec{\theta},\vec{x},\sigma,M_1,I)d\vec{\theta} \tag{195}$$

where $\vec{\theta}$ represents the set of parameters which characterize model M_1. Assume a sufficiently large number of well behaved data. In such a case the likelihood $p(\vec{d}|\vec{\theta},\vec{x},\sigma,M_1,I)$ is a strongly peaked function around some value $\hat{\vec{\theta}}$. The prior in $\vec{\theta}$ varies comparatively slowly and the prior $p(\vec{\theta}|I)$ can be taken out of the integral with $\vec{\theta}$ set to $\hat{\vec{\theta}}$.

$$\begin{aligned} p(\vec{d}|\vec{x},\sigma,M_1) &\approx p(\hat{\vec{\theta}}|M_1)\int d\vec{\theta}\mathcal{L}(\vec{\theta}) \\ &\approx p(\hat{\vec{\theta}}|M_1)\cdot\mathcal{L}(\hat{\vec{\theta}})\cdot(d\theta)^{E_1} \end{aligned} \tag{196}$$

E_1 is the dimension of the parameter space in model M_1. From the normalization of the prior

$$\int p(\vec{\theta}|M_1)d\vec{\theta} = 1 = p(\hat{\vec{\theta}}|M_1)\cdot(\Delta\theta)^{E-1} \tag{197}$$

we obtain an estimate of the prior volume $(\Delta\theta)^{E_1}$ in parameter space. (196) and (197) can of course as well be formulated for a model M_2. The combined result yields the Bayes factor

$$B_{1,2} = \frac{\mathcal{L}(\hat{\vec{\theta}}|M_2)}{\mathcal{L}(\hat{\vec{\theta}}|M_1)}\cdot\frac{(d\theta)^{E_2}}{(d\theta)^{E_1}}\cdot\frac{(\Delta\theta)^{E_1}}{(\Delta\theta)^{E_2}} \tag{198}$$

which simplifies to

$$B_{1,2} = \frac{\mathcal{L}(\hat{\vec{\theta}}|M_2)}{\mathcal{L}(\hat{\vec{\theta}}|M_1)}\cdot\left(\frac{d\theta}{\Delta\theta}\right)^{(E_2-E_1)} \tag{199}$$

The result contains two factors with mutually opposing trends. Let M_2 be the model which contains more parameters, $E_2 > E_1$. Since we have assumed that model M_1 is entirely

contained in model M_2 the likelihood $\mathcal{L}(\hat{\vec{\theta}}|M_2)$ is greater than $\mathcal{L}(\hat{\vec{\theta}}|M_1)$ or at least equal. This follows simply from the fact that the additional parameters contained in M_2 will allow at least a fit as good as with M_1 and nearly always a better one. The likelihood ratio is therefore ≥ 1. Next we investigate the volume factor. Remember $E_2 > E_1$ and $d\theta < \Delta\theta$. This latter assumption relates to a situation where the prior is less informative than the likelihood. The prior knowledge about the parameters encoded in $p(\vec{\theta}|I)$ is then more diffuse than the localization of the likelihood. The parameter space volume ratio $(d\theta/\Delta\theta)^{(E_2-E_1)}$ is therefore smaller than one and penalizes the use of additional parameters. The volume ratio factor is usually called Occam's factor and the net effect of the mutually opposing trends is called Occam's razor. Occam's razor limits the complexity of models since it requires that additional complexity e.g. more parameters in a model must lead to an increase in the likelihood ratio sufficiently strongly such that the decrease of Occam's factor is overcompensated. This is the mathematical formulation of William of Occam's principle that in case that two different explanations apply equally well to a given situation, then the simpler one should be preferred.

C. Hierarchical models

We shall conclude this chapter on Bayesian model comparison with an example for hierarchical model comparison. Consider a set of data as depicted in Fig.9. The set of data $\{d_i\}$ is taken at the set of independent variable $\{x_i\}$. Let ε_i be the error on data point i and $< \varepsilon_i >= 0$, $< \varepsilon_i^2 >= \sigma_i^2$. We shall then attempt to resolve the question whether the data represent a constant or a linear relationship.

$$M_1: \qquad d_i = sx_i + b + \varepsilon_i, \qquad M_2: \qquad d_i = b + \varepsilon_i \tag{200}$$

For comparison of these models we assume that we have no prior preference such that the prior odds in (189) is equal to one. It remains to evaluate the Bayes factor. The marginal likelihood (evidence for model M_1) is

$$p(\vec{d}|\vec{x},\vec{\sigma},M_1) = \int p(b,s|I)\cdot p(\vec{d}|\vec{x},\vec{\sigma},b,s,I)\,db\,ds \tag{201}$$

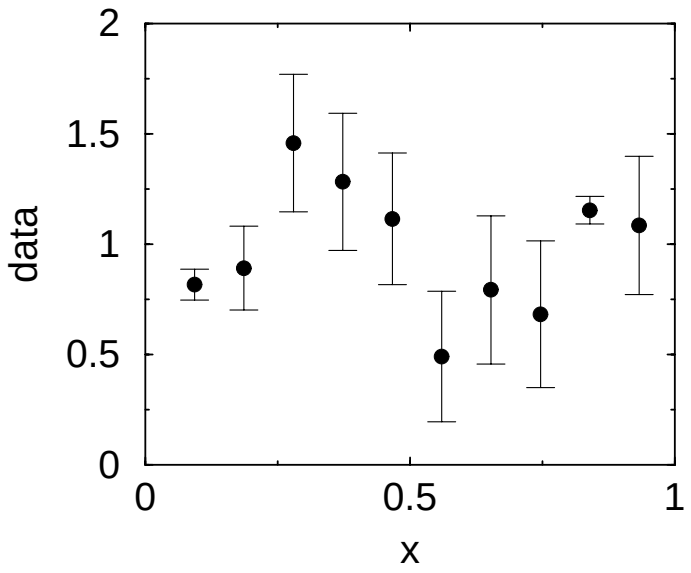

FIG. 9: The data are generated according to $d_i = sx_i + b + \varepsilon_i$ with individually varying noise. Is $s = 0$ or $s \neq 0$ to be expected from this set?

The prior probability $p(b, s|I)$ for model M_1 is already known and given by (144). The likelihood based on the above error assumption is

$$p(\vec{d}|\vec{x}, \vec{\sigma}, b, s, I) = \left(\prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \right) \exp \left\{ -\frac{1}{2} \sum_i (d_i - sx_i - b)^2 / \sigma_i^2 \right\} \tag{202}$$

It is very similar to the previously used form (170,172) where σ was common to all data points. The b-integration is therefore very similar to the previous result (174) and yields

$$p(\vec{d}|\vec{x}, \vec{\sigma}, M_1) = \frac{1}{4B} \cdot \left(\prod_i \frac{1}{\sigma_i \sqrt{2\pi}} \right) \left\{ \frac{2\pi\rho^2}{N} \right\}^{1/2}$$
$$\int \frac{ds}{(1+s^2)^{3/2}} \cdot \exp \left\{ -\frac{N}{2\rho^2} \left(\overline{\Delta D^2} - 2s\overline{\Delta D \Delta x} + s^2 \overline{\Delta x^2} \right) \right\} \tag{203}$$

The average values and the definition of ρ follow the convention (150). The remaining s-integral in (203) cannot be done in closed form. We have previously applied the Gaussian approximation to this type of integral (175). For the present purpose we shall be content with the even simpler peak approximation (196) and take the prior factor out of the integral with $s = s^*$, the maximum of the likelihood

$$s^* = \overline{\Delta D \Delta x} / \overline{\Delta x^2} \tag{204}$$

The remaining integral is then of the standard Gaussian type (22,28) and the evidence becomes

$$p(\vec{d}|\vec{\alpha},\vec{\sigma},M_1) = \frac{1}{4B}\cdot\left(\prod_i \frac{1}{\sigma_i\sqrt{2\pi}}\right)\left\{\frac{2\pi\rho^2}{N}\right\}^{1/2}\cdot\left\{\frac{2\pi\rho^2}{N\overline{\Delta x^2}}\right\}^{1/2}$$
$$\frac{1}{(1+\overline{\Delta D\Delta x}/\overline{\Delta x^2})^{3/2}}\cdot\exp\left\{-\frac{N\,\overline{\Delta D^2}}{2\rho^2}\left(1-\frac{\overline{(\Delta x\Delta D)}^2}{\overline{\Delta x^2}\;\;\overline{\Delta D^2}}\right)\right\} \tag{205}$$

The evidence for model M_2 can be picked from the above calculation (203) before the integration over s is carried out by letting s pass to zero.

$$p(\vec{d}|\vec{x},\vec{\sigma},M_2) = \frac{1}{4B}\left(\prod_i \frac{1}{\sigma_i\sqrt{2\pi}}\right)\left(\frac{2\pi\rho^2}{N}\right)^{1/2}\exp\left\{-\frac{N}{2\rho^2}\cdot\overline{\Delta D^2}\right\} \tag{206}$$

From (205, 206) we derive the odds ratio O

$$O = \frac{1}{2}\left(\frac{2\pi\rho^2}{N\overline{\Delta x^2}}\right)^{1/2}\frac{1}{(1+\overline{\Delta D\Delta x}/\overline{\Delta x^2})^{3/2}}\cdot\exp\left\{\frac{N}{2\rho^2}\cdot\frac{\overline{(\Delta x\Delta D)}^2}{\overline{\Delta x^2}}\right\} \tag{207}$$

Recall that $O > 1$ favours a linear relationship between $\vec{x}$ and $\vec{d}$ while $O < 1$ favours a constant. Of critical importance for the value of O is the Occam factor

$$(1+\overline{\Delta x\Delta D}/\overline{\Delta x^2})^{-3/2} \tag{208}$$

which follows from the prior in s. Suppose, that we had taken a flat prior in s within some reasonable limits $p(s) = 1/\Delta S$. The odds ratio decreases then linearly with ΔS and would eventually favour the simpler constant model M_2 even if the data show a clear linear trend. In Bayesian model comparison one does not get away with pretending innocence. On the contrary, one has to specify the prior knowledge as precisely as possible. In particular improper priors (110), (124) are of no use at all in comparison of models containing different numbers of parameters.

The odds for a linear versus a constant model for the data in Fig.9 favour the linear model with an odds of +3.65. Note that the data were generated with a slope of +0.3. The straight forward least squares result neglecting the different errors yields a slope of -0.104 while the correctly weighted data yield a slope of 0.365. This shows the importance of correct account for individual errors. In fact, the preference for a linear model to explain the data in Fig.9 rests nearly totally on the two points with very small error. Intuition is not a good guide

in this case, since it appears to be based on the scatter of the data only. Our brain cannot spontaneously deal correctly with confidence margins.

Here ends the introduction to Bayesian analysis in physics. The examples presented were chosen as realistically as possible and yet they leave the impression of being slightly artificial. The choice was primarily dictated by the involved computational complexity. This rises strongly as one leaves one- or twodimensional problems. A forthcoming second series of lectures will therefore start with techniques for efficient evaluation of multidimensional integrals as a prerequisite to treat problems as they arise in an experimentalist's everyday work.