Linear Regression With Errors in Both Variables

T. Ihn

21. November 2016

Background information and model: We measure data points $(x_{\mathrm{m}i}, y_{\mathrm{m}i})$, where both quantities are independently measured. For example, $x_{\mathrm{m}i}$ could be the measured value of a magnetic field and $y_{\mathrm{m}i}$ would be the the corresponding Hall voltage. The special situation considered here is that both quantities have an additive error, i.e.

$$\begin{aligned} x_{\mathrm{m}i} &= x_i + \epsilon_i \\ y_{\mathrm{m}i} &= y_i + \eta_i \end{aligned}$$

The errors ϵ_i and η_i are assumed to be statistically independent.

We consider the case where a linear functional relationship exists between the exact values y_i and x_i. Therefore,

$$y_i = \alpha x_i + \beta \tag{1}$$

is considered to be a valid model for the measured data with parameters α and β independent of the measured data point. The probability distributions for ϵ_i and η_i are given by normal distributions

$$\begin{aligned} \mathrm{pdf}(\epsilon_i|\sigma_x)d\epsilon_i &= \mathcal{N}(\epsilon_i; 0, \sigma_x)d\epsilon_i \\ \mathrm{pdf}(\eta_i|\sigma_y)d\eta_i &= \mathcal{N}(\eta_i; 0, \sigma_y)d\eta_i. \end{aligned}$$

The error bars σ_x and σ_y are assumed to be known, and to be the same for all data points. An example dataset is shown in Fig. 1.

General difficulty of this problem. We compare this problem to the standard linear regression case with the N measured data points $(x_{\mathrm{m}i}, y_{\mathrm{m}i})$, where all the $x_{\mathrm{m}i} \equiv x_i$ are known exactly. In this case there remain N unknown exact values y_i. Given the linear relation (1), knowledge of the two unknown parameters α and β would allow us to immediately determine all the unknown exact y_i values from the $\{x_i\}$. The linear fitting function therefore *reduces* the N unknown values $\{y_i\}$ to only two unknown parameters. We can extract estimates and uncertainties for these two paramters α and β from $N > 2$ data points.

Now we see the general difficulty of having errors in both variables. Having N measured data points $(x_{\mathrm{m}i}, y_{\mathrm{m}i})$ we start with $2N$ unknowns, i.e., the N pairs of exact values $\{(x_i, y_i)\}$.

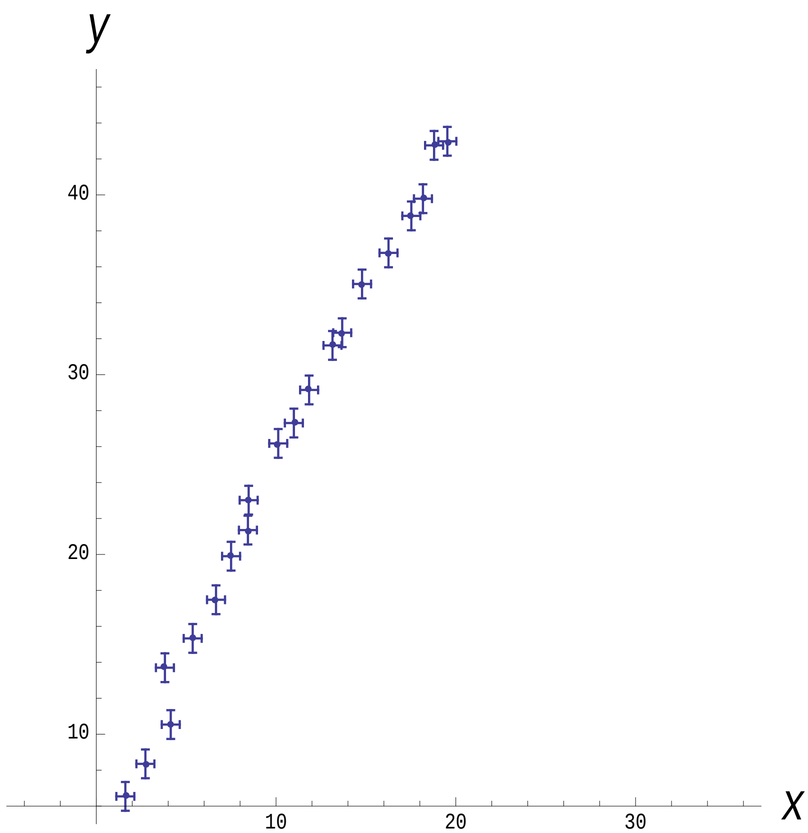

Abbildung 1: Data with errors in both variables. The goal is to find a straight line fit taking the errors in both variables into account. For this data, $\sigma_x = 0.5$ and $\sigma_y = 0.8$ are known.

Using the linear relation between the x_i and the y_i gives N equations, but two additional parameters, leading to a net reduction of the number of unknown parameters to $N + 2$, e.g. the N values $\{x_i\}$ and the two paramters α and β. However, we still have only N data points, i.e., the system is still underdetermined (rather than overdetermined). This means that we cannot get a reasonable linear fit and extract the parameters α and β without supplying *additional information* about some of the $N+2$ unknown parameters. It is clear that a similar problem would arise, if the errors σ_x and σ_y were unknown as well. This would further increase the number of unknown parameters to $N + 4$.

A different view of the same problem. We now follow Gull[1] in looking at the same problem from a different angle, which highlights the symmetry between the x- and y-values.

[1]Stephen F. Gull, *Bayesian Data Analysis: Straight-line Fitting* in: Maximum Entropy and Bayesian Methods, J. Skilling (ed.), (Kluwer Academic Publishers, Dordrecht, 1988), pp. 53–74.

Suppose we consider dimensionless quantities

$$\begin{aligned} x_i' &= \frac{x_i - x_0}{r_x} \\ y_i' &= \frac{y_i - y_0}{r_y}, \end{aligned}$$

where x_0 and y_0 are location parameters and $r_x > 0$ and $r_y > 0$ are scale parameters. The functional relationship (1) between x_i' and y_i' can then be written as

$$r_y y_i' + y_0 = \alpha(r_x x_i' + x_0) + \beta.$$

Identifying

$$\alpha = \pm\frac{r_y}{r_x} \quad \text{and} \quad \beta = y_0 - \alpha x_0$$

we find the line with slope one (or minus one)

$$y_i' = \pm x_i'.$$

We may therefore say, instead of looking for a slope parameter α and an intersection parameter β, we look for location parameters x_0, y_0 and scale parameters r_x, r_y that would transform our data to a line of slope one (or -1, if $\alpha < 0$) extending in x_i' and y_i' symmetrically around the origin. The error variables ϵ_i and η_i transform according to

$$\epsilon_i' = \frac{\epsilon_i}{r_x} \quad \text{and} \quad \eta_i' = \frac{\eta_i}{r_y}.$$

Correspondingly, their distribution functions are

$$\mathrm{pdf}(\epsilon_i') = \mathcal{N}(\epsilon_i'; 0, \sigma_x') \quad \text{and} \quad \mathrm{pdf}(\eta_i') = \mathcal{N}(\eta_i'; 0, \sigma_y')$$

with

$$\sigma_x' = \frac{\sigma_x}{r_x} \quad \text{and} \quad \sigma_y' = \frac{\sigma_y}{r_y}$$

and the measured data are

$$\frac{x_{\mathrm{m}i} - x_0}{r_x} = x_i' + \epsilon_i' \quad \text{and} \quad \frac{y_{\mathrm{m}i} - y_0}{r_y} = y_i' + \eta_i'.$$

The likelihood function. We further assume the two uncertainties η_i' and ϵ_i' to be statistically independent and obtain the joint distribution

$$\mathrm{pdf}(\epsilon_i', \eta_i' | \sigma_x', \sigma_y') d\epsilon_i' d\eta_i' = \mathcal{N}(\epsilon_i'; 0, \sigma_x')\mathcal{N}(\eta_i'; 0, \sigma_y') d\epsilon_i' d\eta_i'.$$

As a result we have the samping distribution for N data points

$$\begin{aligned}
&\mathrm{pdf}(\{x_{\mathrm{m}i}, y_{\mathrm{m}i}\}|\{x_i\}, x_0, y_0, r_x, r_y, \sigma'_x, \sigma'_y) d^N x_{\mathrm{m}} d^N y_{\mathrm{m}} \\
&\qquad = \frac{d^N x_{\mathrm{m}i} d^N y_{\mathrm{m}i}}{(r_x r_y)^N} \prod_{i=1}^{N} \mathcal{N}\left(\frac{x_{\mathrm{m}i} - x_0}{r_x}; x'_i, \sigma'_x\right) \mathcal{N}\left(\frac{y_{\mathrm{m}i} - y_0}{r_y}; x'_i, \sigma'_y\right) \\
&= \frac{d^N x_{\mathrm{m}i} d^N y_{\mathrm{m}i}}{(2\pi\sigma'_x\sigma'_y r_x r_y)^N} \exp\left[-\frac{1}{2}\left(\frac{\sum_{i=1}^{N}(x_{\mathrm{m}i} - x_i)^2}{{\sigma_x}^2} + \frac{\sum_{i=1}^{N}(y_{\mathrm{m}i} - y_0 - r_y(x_i - x_0)/r_x)^2\rangle}{{\sigma_y}^2}\right)\right].
\end{aligned}$$

This likelihood function contains the N parameters x_i as so-called *nuisance parameters.*[2] In addition, we have replaced the original two parameters α and β by *four* new parameters x_0, y_0, r_x, and r_y, which may seem a bit awkward at a first glance. However, we will see below that this allows a formulation of the problem that is symmetric under the exchange of x and y, a symmetry that naturally appears in the problem, because it is usually arbitrary, which of the two measured quantities we call x and which y.

Prior distributions for x_0, y_0, r_x, and r_y. In order to make further progress, we need a prior distribution function for the $N + 4$ parameters

$$\mathrm{pdf}(\{x_i\}, x_0, y_0, r_x, r_y) = \mathrm{pdf}(x_0, y_0, r_x, r_y)\mathrm{pdf}(\{x_i\}|x_0, y_0, r_x, r_y).$$

We choose

$$\mathrm{pdf}(x_0, y_0, \ln r_x, \ln r_y) \propto dx_0 dy_0 d(\ln r_x) d(\ln r_y),$$

which is a completely uninformative prior for the location parameters x_0 and y_0, and for the scale parameters $r_x > 0$ and $r_y > 0$. For the latter, a uniform distribution in the logarithm ensures that these quantities are positive. Such a prior is called Jeffreys prior.[3]

Prior distribution for the $\{x_i\}$. There remains a prior to be found for the unknown $\{x_i\}$. Using a completely uninformative (constant) prior for these variables would leave us with an unsolvable underdetermined problem, as discussed above. We therefore have to provide additional information here about the $\{x_i\}$. We will therefore assume that all the x_i are with large probability within the range $\pm r_x$ around x_0. We see that this leaves great freedom for the x_i to vary within the range, where their values are expected. This is the additional information that we supply to return to an overdetermined problem in the end.

In order to implement this thought, we use the product of gaussians

$$\mathrm{pdf}(\{x_i\}|x_0, y_0, r_x, r_y) = \prod_{i=1}^{N} \mathcal{N}(x_i; x_0, r_x) = \frac{1}{(2\pi r_x^2)^{N/2}} \exp\left[-\frac{1}{2}\frac{\sum_{i=1}^{N}(x_i - x_0)^2}{r_x^2}\right]$$

[2] Nuisance parameters are parameters of a model that are not of immediate interest in the analysis. Here we are aiming at the determination of α and β. The true positions x_i are not of interest to us.

[3] Sir Harold Jeffreys suggested the use of this prior for (positive) scale parameters in his book *Theory of Probability.*

Note that this prior is completely symmetric in x and y, since our choice of parameters ensures that $(x_i - x_0)/r_x = (y_i - y_0)/r_y$. It therefore conforms with our aim of a formulation of the problem symmetric under the exchange of x and y.

It is a general property of models with nuisance parameters, that a prior distribution needs to be specified, which will influence the final result. This prior allows us to marginalize the nuisance parameters later on. The choice of a gaussian prior in our specific case is less a matter of necessity, but of convenience. We will see that it later allows an analytic marginalization of the nuisance parameters.

Joint distribution. We can now write down the joint distribution function for data and parameters as

$$\begin{aligned}&\mathrm{pdf}(\{x_{\mathrm{m}i}, y_{\mathrm{m}i}\}, \{x_i\}, x_0, y_0, \ln r_x, \ln r_y | \sigma_x, \sigma_y)\\ &\qquad = \frac{1}{(8\pi^3 \sigma_x^2 \sigma_y^2 r_x^2)^{N/2}}\\ &\qquad \times \exp\left[-\frac{1}{2} \sum_{i=1}^{N} \left(\frac{(x_{\mathrm{m}i} - x_i)^2}{\sigma_x^2} + \frac{(y_{\mathrm{m}i} - y_0 - r_y(x_i - x_0)/r_x)^2}{\sigma_y^2} + \frac{(x_i - x_0)^2}{r_x^2} \right) \right].\end{aligned}$$

Integrating out (marginalizing) the nuisance parameters x_i. In the next step we integrate out the nuisance parameters x_i. This multidimensional integral separates into N integrals of the form

$$\begin{aligned}&r_x \int dx_i' \exp\left[-\frac{1}{2} \left(\frac{(x_{\mathrm{m}i}' - x_i')^2}{{\sigma_x'}^2} + \frac{(y_{\mathrm{m}i}' - x_i')^2}{{\sigma_y'}^2} + {x_i'}^2 \right) \right]\\ &\qquad = \frac{\sqrt{2\pi}\sigma_x'\sigma_y' r_x}{\sqrt{{\sigma_x'}^2 + {\sigma_y'}^2 + {\sigma_x'}^2{\sigma_y'}^2}} \times \exp\left[-\frac{(1 + {\sigma_y'}^2){x_{\mathrm{m}i}'}^2 - 2x_{\mathrm{m}i}'y_{\mathrm{m}i}' + (1 + {\sigma_x'}^2){y_{\mathrm{m}i}'}^2}{2({\sigma_x'}^2 + {\sigma_y'}^2 + {\sigma_x'}^2{\sigma_y'}^2)} \right],\end{aligned}$$

where the exponent is a quadratic form of x_0 and y_0. The result of the N-fold integration is therefore the joint distribution for the data and the 4 remaining parameters

$$\begin{aligned}&\mathrm{pdf}(\{x_{\mathrm{m}i}, y_{\mathrm{m}i}\}, x_0, y_0, \ln r_x, \ln r_y | \sigma_x, \sigma_y)\\ &\qquad = \left(\frac{1}{4\pi^2 \left(r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2 \right)} \right)^{N/2}\\ &\times \exp\left[-\frac{\sum_{i=1}^{N} \left[(r_y^2 + {\sigma_y}^2)(x_{\mathrm{m}i} - x_0)^2 - 2r_x r_y (x_{\mathrm{m}i} - x_0)(y_{\mathrm{m}i} - y_0) + (r_x^2 + {\sigma_x}^2)(y_{\mathrm{m}i} - y_0)^2 \right]}{2(r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2)} \right],\end{aligned}$$

which is a bivariate gaussian distribution for x_0 and y_0. Note also the symmetry of the joint distribution with respect of an exchange of x and y.

Sufficient statistics. The sum in the numerator of the exponent can be transformed into sample averages giving

$$\begin{aligned}
&\mathrm{pdf}(\{x_{\mathrm{m}i}, y_{\mathrm{m}i}\}, x_0, y_0, \ln r_x, \ln r_y | \sigma_x, \sigma_y) \\
&\qquad = \left(\frac{1}{4\pi^2 \left(r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2 \right)} \right)^{N/2} \\
&\times \exp\left[-\frac{N}{2} \frac{(r_y^2 + {\sigma_y}^2)(x_0 - \overline{x_{\mathrm{m}i}})^2 - 2 r_x r_y (x_0 - \overline{x_{\mathrm{m}i}})(y_0 - \overline{y_{\mathrm{m}i}}) + (r_x^2 + {\sigma_x}^2)(y_0 - \overline{y_{\mathrm{m}i}})^2}{r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2} \right] \\
&\times \exp\left[-\frac{N}{2} \frac{(r_y^2 + {\sigma_y}^2)\mathrm{Var}(x_{\mathrm{m}i}) - 2 r_x r_y \sqrt{\mathrm{Var}(x_{\mathrm{m}i})\mathrm{Var}(y_{\mathrm{m}i})}\rho + (r_x^2 + {\sigma_x}^2)\mathrm{Var}(y_{\mathrm{m}i})}{r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2} \right].
\end{aligned}$$

We see that the quantities $\overline{x_{\mathrm{m}i}}$, $\overline{y_{\mathrm{m}i}}$, $\mathrm{Var}(x_{\mathrm{m}i})$, $\mathrm{Var}(y_{\mathrm{m}i})$, and ρ are a sufficient statistic for the problem, like in standard linear regression where errors are only in y.

We note here that the exponent of the first exponential factor can be expressed as

$$-\frac{N}{2} \begin{pmatrix} x_0 - \overline{x_{\mathrm{m}i}} \\ y_0 - \overline{y_{\mathrm{m}i}} \end{pmatrix} \underbrace{\frac{1}{r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2} \begin{pmatrix} r_y^2 + \sigma_y^2 & -r_x r_y \\ -r_x r_y & r_x^2 + \sigma_x^2 \end{pmatrix}}_{:=M} \begin{pmatrix} x_0 - \overline{x_{\mathrm{m}i}} \\ y_0 - \overline{y_{\mathrm{m}i}} \end{pmatrix},$$

where $\det(M) = 1$.

Posterior distribution and estimates of the shift parameters. The posterior distribution for x_0, y_0, r_x, r_y given the data is then

$$\begin{aligned}
&\mathrm{pdf}(x_0, y_0, \ln r_x, \ln r_y | \{x_{\mathrm{m}i}, y_{\mathrm{m}i}\}, \sigma_x, \sigma_y) \\
&\qquad \propto \left(r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2 \right)^{-N/2} \\
&\times \exp\left[-\frac{N}{2} \frac{(r_y^2 + {\sigma_y}^2)(x_0 - \overline{x_{\mathrm{m}i}})^2 - 2 r_x r_y (x_0 - \overline{x_{\mathrm{m}i}})(y_0 - \overline{y_{\mathrm{m}i}}) + (r_x^2 + {\sigma_x}^2)(y_0 - \overline{y_{\mathrm{m}i}})^2}{r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2} \right] \\
&\times \exp\left[-\frac{N}{2} \frac{(r_y^2 + {\sigma_y}^2)\mathrm{Var}(x_{\mathrm{m}i}) - 2 r_x r_y \sqrt{\mathrm{Var}(x_{\mathrm{m}i})\mathrm{Var}(y_{\mathrm{m}i})}\rho + (r_x^2 + {\sigma_x}^2)\mathrm{Var}(y_{\mathrm{m}i})}{r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2} \right].
\end{aligned}$$

From this posterior we find the estimates for x_0 and y_0 with their uncertainties

$$\begin{aligned}
\hat{x}_0 = \langle x_0 \rangle = \overline{x_{\mathrm{m}i}} \qquad &\text{and} \qquad \hat{y}_0 = \langle y_0 \rangle = \overline{y_{\mathrm{m}i}} \\
\langle \Delta x_0^2 \rangle = \frac{\sigma_x^2 + r_x^2}{N} \qquad &\text{and} \qquad \langle \Delta y_0^2 \rangle = \frac{\sigma_y^2 + r_y^2}{N} \\
\langle \Delta x_0 \Delta y_0 \rangle &= \tfrac{r_x r_y}{N}
\end{aligned}$$

Note that the estimates taken to be the mean values of x_0 and y_0 calculated with the posterior distribution are at the same time maximizing the posterior distribution for any values of r_x and r_y. These estimates allow us to plot the shifted data as shown in Fig. 2.

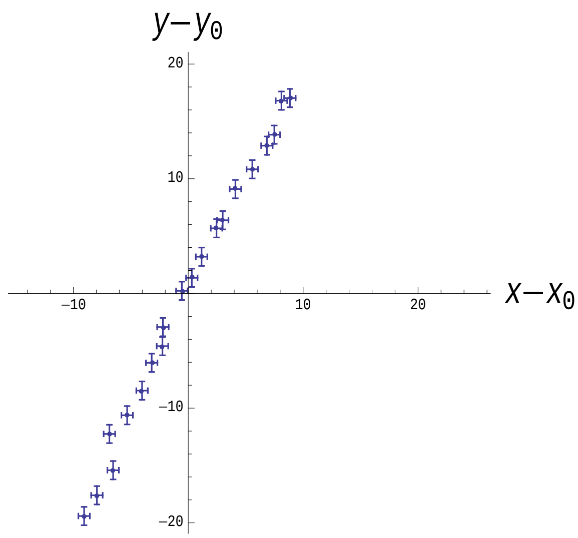

Abbildung 2: Data shifted by $x_0 = \overline{x_{\mathrm{m}i}} = 10.68$ and $y_0 = \overline{y_{\mathrm{m}i}} = 25.95$.

Estimating the scaling parameters. Integrating out x_0 and y_0 from the posterior distribution gives

$$\begin{aligned} &\mathrm{pdf}(\ln r_x, \ln r_y | \{x_{\mathrm{m}i}, y_{\mathrm{m}i}\}, \sigma_x, \sigma_y) \\ &\qquad \propto \left(r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2\right)^{-N/2} \\ &\quad \times \exp\left[-\frac{N}{2} \frac{(r_y^2 + {\sigma_y}^2)\mathrm{Var}(x_{\mathrm{m}i}) - 2 r_x r_y \sqrt{\mathrm{Var}(x_{\mathrm{m}i})\mathrm{Var}(y_{\mathrm{m}i})}\rho + (r_x^2 + {\sigma_x}^2)\mathrm{Var}(y_{\mathrm{m}i})}{r_y^2 {\sigma_x}^2 + r_x^2 {\sigma_y}^2 + {\sigma_x}^2 {\sigma_y}^2}\right]. \end{aligned}$$

We see here that $\mathrm{Var}(x_{\mathrm{m}i})$ and $\mathrm{Var}(y_{\mathrm{m}i})$ appear as natural scales of the problem. If we introduce

$$r'_x = \frac{r_x}{\sqrt{\mathrm{Var}(x_{\mathrm{m}i})}}, \quad r'_y = \frac{r_y}{\sqrt{\mathrm{Var}(y_{\mathrm{m}i})}}, \quad \sigma'_x = \frac{\sigma_x}{\sqrt{\mathrm{Var}(x_{\mathrm{m}i})}}, \quad \sigma'_y = \frac{\sigma_y}{\sqrt{\mathrm{Var}(y_{\mathrm{m}i})}},$$

we obtain

$$\begin{aligned} &\mathrm{pdf}(\ln r'_x, \ln r'_y | \{x_{\mathrm{m}i}, y_{\mathrm{m}i}\}, \sigma'_x, \sigma'_y) \\ &\qquad \propto \left({r'_y}^2 {\sigma'_x}^2 + {r'_x}^2 {\sigma'_y}^2 + {\sigma'_x}^2 {\sigma'_y}^2\right)^{-N/2} \\ &\qquad\qquad \times \exp\left[-\frac{N}{2} \frac{({r'_y}^2 + {\sigma'_y}^2) - 2 r'_x r'_y \rho + ({r'_x}^2 + {\sigma'_x}^2)}{{r'_y}^2 {\sigma'_x}^2 + {r'_x}^2 {\sigma'_y}^2 + {\sigma'_x}^2 {\sigma'_y}^2}\right]. \end{aligned}$$

We may now estimate the parameters r'_x and r'_y from the negative logarithm of this posterior function numerically. The example for our data is shown in Fig. 3. We may check our result

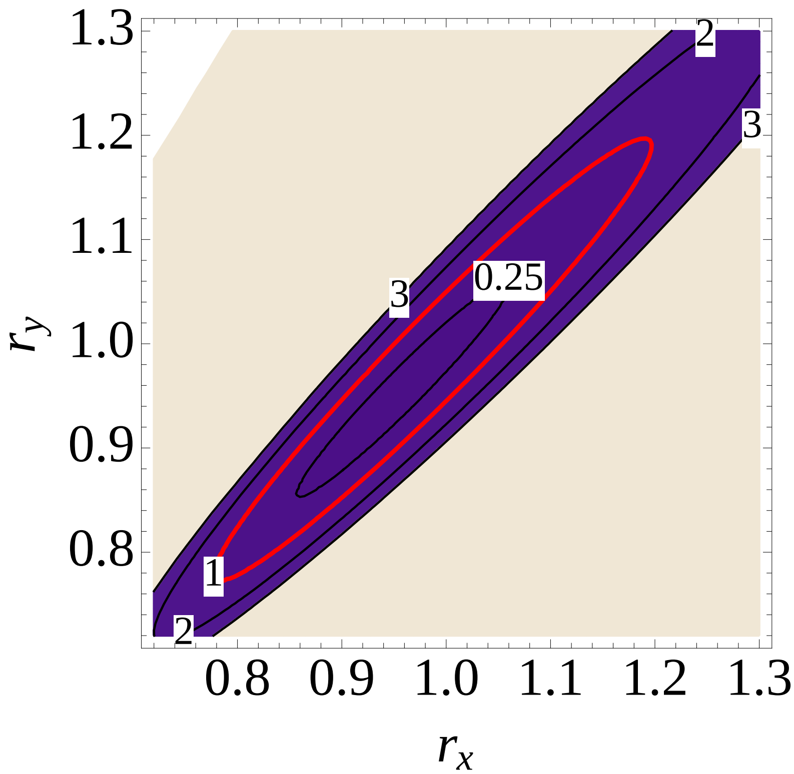

Abbildung 3: Logarithm of the posterior distribution for r_x and r_y. The minimum of this function is at $r'_x = r'_y = 0.95$. The red contour line at the value 1 may be used for estimating the standard errors in these quantities.

by calculating the corresponding estimates for r_x and r_y. The data can then be plotted in the scaled coordinates as shown in Fig. 4.

Estimating the slope α of the data. Changing variables to $\alpha' = r'_y/r'_x$ and $R' = \sqrt{r'_x r'_y}$ gives

$$r'_x = \frac{R'}{\sqrt{\alpha'}} \quad \text{and} \quad r'_y = R'\sqrt{\alpha'}.$$

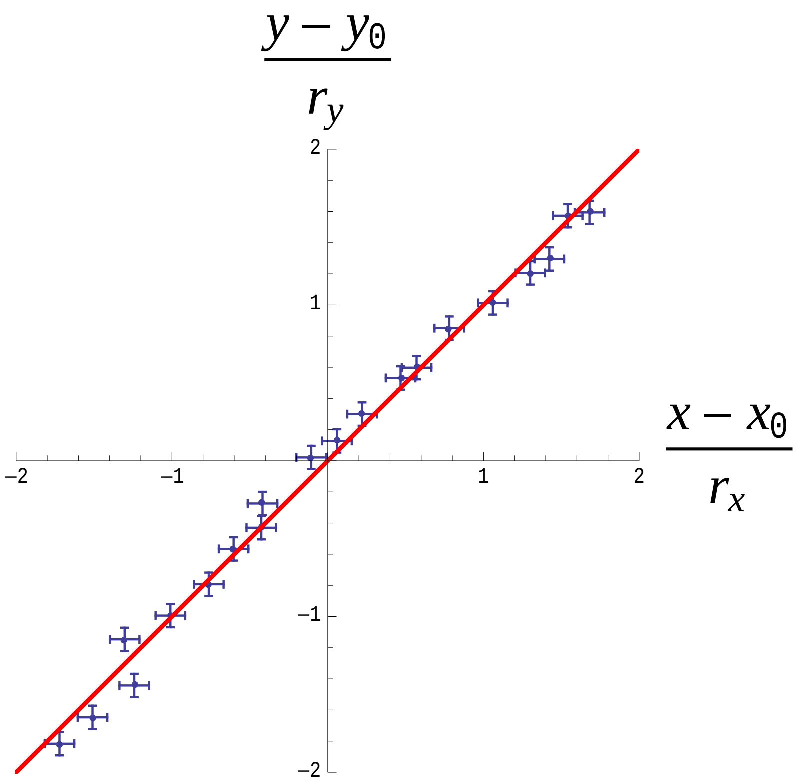

Abbildung 4: The original data shifted by (x_0, y_0) and scaled by the estimated r_x and r_y. The red line corresponds to the diagonal along which the data are expected to be scattered.

It leads to

$$\begin{aligned}
&\mathrm{pdf}(\ln\alpha', \ln R' | \{x_{\mathrm{m}i}, y_{\mathrm{m}i}\}, \sigma'_x, \sigma'_y) \\
&\qquad \propto \left(R'^2\alpha'\sigma'^2_x + R'^2\sigma'^2_y/\alpha' + \sigma'^2_x\sigma'^2_y\right)^{-N/2} \\
&\qquad\qquad \times \exp\left[-\frac{N}{2}\frac{R'^2\alpha'^2 - 2R'^2\rho\alpha' + (\sigma'^2_x + \sigma'^2_y)\alpha' + R'^2}{R'^2\alpha'^2\sigma'^2_x + (\sigma'^2_x\sigma'^2_y)\alpha' + R'^2\sigma'^2_y}\right].
\end{aligned}$$

The minimum of the negative logarithm of this function shown in Fig. 5 is a direct way to estimate α' and its uncertainty.

Estimating the intercept parameter β. The quantity β may now be estimated via the relation $\beta = y_0 - \alpha x_0$ to be

$$\beta = \overline{y_{\mathrm{m}i}} - \alpha\overline{x_{\mathrm{m}i}}.$$

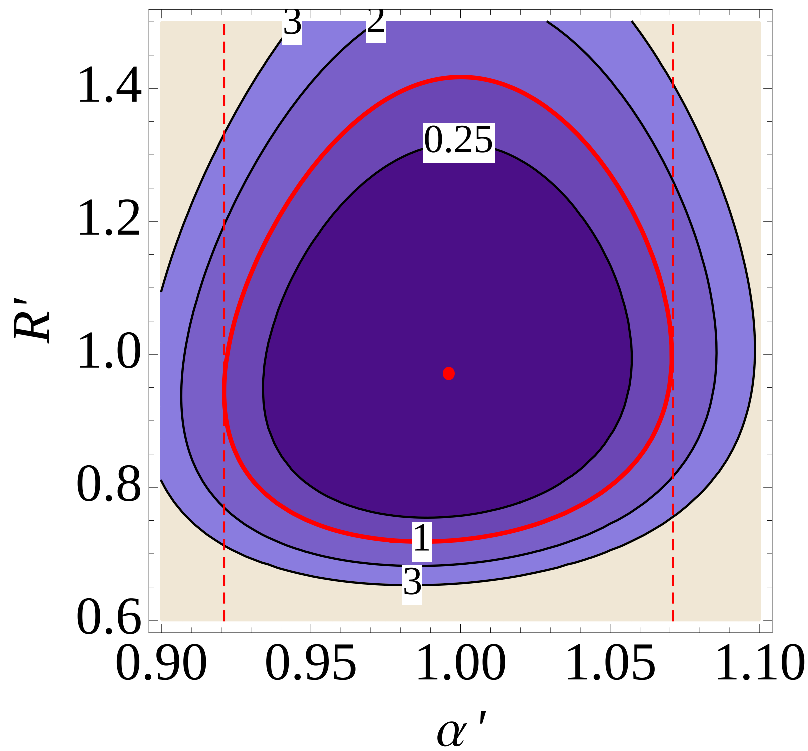

Abbildung 5: Negative logarithm of the posterior distribution for α' and R'. The minimum in this figure is found at $\alpha' = 0.996$ and $R' = 0.971$. Together with the data variance $\mathrm{Var}(x_{\mathrm{m}i}) = 30.57$ and $\mathrm{Var}(y_{\mathrm{m}i}) = 127.00$ this gives an estimate of the slope $\alpha = 2.03$. The red contour line may be used to estimate the errors in the two quantities graphically. We find from the two vertical dashed lines $\alpha' = 0.996 \pm 0.075$ translating into $\alpha = 2.03 \pm 0.15$.

In principle, the uncertainty of the b-estimate would need to be calculated from the posterior distribution for x_0, y_0 and α. The integration over x_0 and y_0 can be performed analytically and gives

$$\langle \Delta\beta^2 \rangle_{x_0,y_0} = \frac{\sigma_y^2 + \alpha^2 \sigma_x^2}{N}.$$

As a shortcut we may use, instead of the numerical integration, using Gauss' error propagation law

$$\langle \Delta\beta^2 \rangle = \frac{\sigma_y^2 + \alpha_{\mathrm{est}}^2 \sigma_x^2}{N} + \overline{x_{\mathrm{m}i}}^2 \langle \Delta\alpha^2 \rangle.$$

The final result of the fit. Eventually we show the final result of the original data, together with the fitted line determined above in Fig. 6.

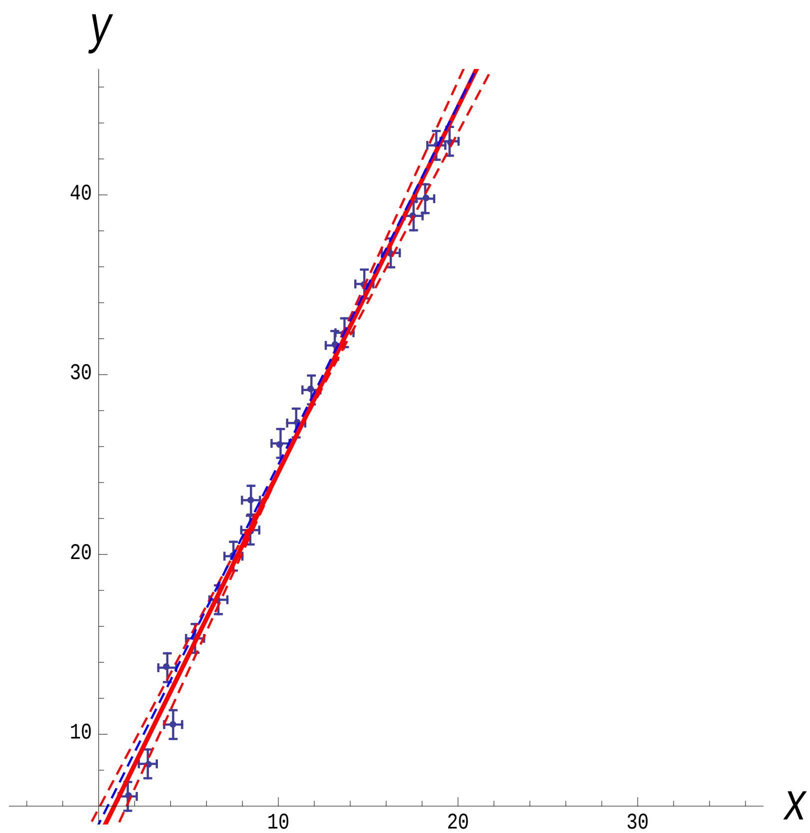

Abbildung 6: Final result of the fit to data with errors in x and y. The solid red line represents the fit with $\alpha = 2.03$ and $\beta = 4.3$. The error in β is $\Delta\beta = 1.6$. The red dashed lines have slopes $\alpha \pm \Delta\alpha = 2.03 \pm 0.15$ and $\beta = 4.3$. The blue dashed line represents the 'true' curve $y = 2x + 5$ from which the data have been generated.

Python code realizing the minimization.

```
from csv import reader

import numpy as np
import matplotlib as mpl
import matplotlib.pyplot as plt

from scipy.optimize import fmin
```

```
#
# read data from a csv-file
#
readerout = reader(open("errorxydata.csv","rb"),delimiter=',');
x = list(readerout);
data = np.array(x).astype('float')
xi = data[:,0]
yi = data[:,1]

# Error bars
sigmax = 0.5
sigmay = 0.8

#
# Statistics of the data
#

# number of data points
nn = len(data)
print('\nStatistics of the data:\nN = %1u data points'% (nn))

# mean of x-values
Meanx = np.mean(xi)
print('<x> = %1.2f'% (Meanx))

# mean of y-values
Meany = np.mean(yi)
print('<y> = %1.2f'% (Meany))

# Variance of x-values
Varx = np.var(xi)
print('Var(x) = %1.2f'% (Varx))

# Variance of y-values
Vary = np.var(yi)
print('Var(y) = %1.2f'% (Vary))

# Empirical correlation coefficient
rho = np.corrcoef(xi.T,yi.T)[0,1]
print('rho = %1.4f\n'% (rho))
```

```
# Define the -log-posterior for a and R
def posteriorpdf(parms,sx,sy,rho,n):
    a = parms[0]
    R = parms[1]
    num = R*R*a*a - 2*R*R*rho*a + (sx*sx+sy*sy)*a + R*R
    denom = sx*sx*R*R*a*a + sx*sx*sy*sy*a + R*R*sy*sy
    p = num/denom
    q = sx*sx*R*R*a+sy*sy*R*R/a+sx*sx*sy*sy
    return np.log(R*a) + (n/2.)*np.log(q) + (n/2.)*p

# Minimization of the -log-posterior
print('Estimating scaled parameters:')
parms,Qmin,_,_,_ = fmin(posteriorpdf,[1.,1.],args=(sigmax/np.sqrt(Varx),sigmay/np.sqrt(Vary
aa = parms[0]
RR = parms[1]

# plot the -log-posterior
a = np.arange(0.93,1.07,0.001)
R = np.arange(0.7,1.3,0.01)
Q = np.zeros(shape=(len(R),len(a)))
X,Y = np.meshgrid(a,R)
for n in range(0,len(a)):
    for m in range(0,len(R)):
        Q[m,n] = posteriorpdf([a[n],R[m]],sigmax/np.sqrt(Varx),sigmay/np.sqrt(Vary),rho,nn)


plt.figure(
    num=2,
    figsize=(7,5.08),
    dpi=80,
    facecolor='white')
mpl.rcParams['axes.linewidth']=2
plt.axes([0.15,0.18,0.82,0.70],
    axisbg='lightgray')
CS = plt.contourf(X,Y,Q,100)
# We draw two contour lines into the plot, the smaller of which is suitable for reading the
CS2 = plt.contour(CS,levels=[0.2,1,2,3],colors=('w','r','w','w'),linewidths=(2,),hold='on')
plt.clabel(CS2, inline=1, fmt='%1.1f', fontsize=16)
plt.plot(aa,RR,'wo')
plt.title('-log-posterior',fontsize=24)
```

```
plt.xlabel(r'$a\sqrt{\mathrm{Var}(x)/\mathrm{Var}(y)}$',
    fontsize=24)
plt.ylabel(r'$R/(\mathrm{Var}(x)\mathrm{Var}(y))^{1/4}$',
    fontsize=24)
cbar = plt.colorbar(CS)
cbar.ax.set_ylabel(r'$Q(a,R)$',fontsize=24)
plt.show()

# We determine the standard error of the parameters from the contour line
cc = CS2.collections[1].get_paths()[0]
CC = cc.vertices
sam = aa-min(CC[:,0])
sap = max(CC[:,0])-aa
sRm = RR-min(CC[:,1])
sRp = max(CC[:,1])-RR

print('Scaled parameters:')
print('a\' = %1.3f +%1.3f/-%1.3f'% (aa,sap,sam))
print('R\' = %1.2f +%1.2f/-%1.2f\n'% (RR,sRp,sRm))

# Calculating the slope and the intercept parameter and their errors
print('Estimated parameters:')
aest = aa*np.sqrt(Vary/Varx)
sam1 = sam*np.sqrt(Vary/Varx)
sap1 = sap*np.sqrt(Vary/Varx)
print('a = %1.3f +%1.3f/-%1.3f'% (aest,sap1,sam1))

best = Meany - Meanx*aest
bstd = np.sqrt( (sigmay*sigmay+sigmax*sigmax*aest)/nn + Meanx*Meanx*sam1*sap1)
print('b = %1.2f +/-%1.2f'% (best,bstd))

#
# Plot data with error bars and fitted line
#
x = np.arange(min(xi),max(xi),0.1)
y = aest*x + best
plt.figure(
    num=1,
    figsize=(7,5.08),
    dpi=80,
    facecolor='white')
```

```
mpl.rcParams['axes.linewidth']=2
plt.axes([0.15,0.18,0.82,0.70],
    axisbg='lightgray')
plt.errorbar(xi, yi, xerr=sigmax, yerr=sigmay, fmt='o')
plt.plot(x,y,'r',linewidth=2)
plt.axis([0,21,5,45])
plt.xticks([0,10,20],
    fontsize=24,
    fontname='Times New Roman')
plt.yticks([10,20,30,40],
    fontsize=24,
    fontname='Times New Roman')
plt.tick_params(width=2,length=10)
plt.xlabel(r'$x$',
    fontsize=24)
plt.ylabel(r'$y$',
    fontsize=24)
plt.title('Data with errors in x and y',fontsize=24)
plt.text(2,40,r"$y = a x + b$",
    fontsize=24,
    fontname='Times New Roman')
plt.text(2,35,r"$a = 2.039\pm$",
    fontsize=24,
    fontname='Times New Roman')
plt.text(8,36.8,r"$0.076$",
    fontsize=24,
    fontname='Times New Roman')
plt.text(8,33.2,r"$0.073$",
    fontsize=24,
    fontname='Times New Roman')
plt.text(2,30,r"$b = 4.17\pm 0.83$",
    fontsize=24,
    fontname='Times New Roman')

plt.show()
```
