9 Straight-line fitting with errors in both variables

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.

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

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

$$y_i = \alpha x_i + \beta \tag{58}$$

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

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

An example dataset is shown in Fig. 52.

A different view of the same problem. We now follow Gull[7] in looking at the same problem from a different angle, which highlights the symmetry between the x- and y-values. 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 (58) 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

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

[7] 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.

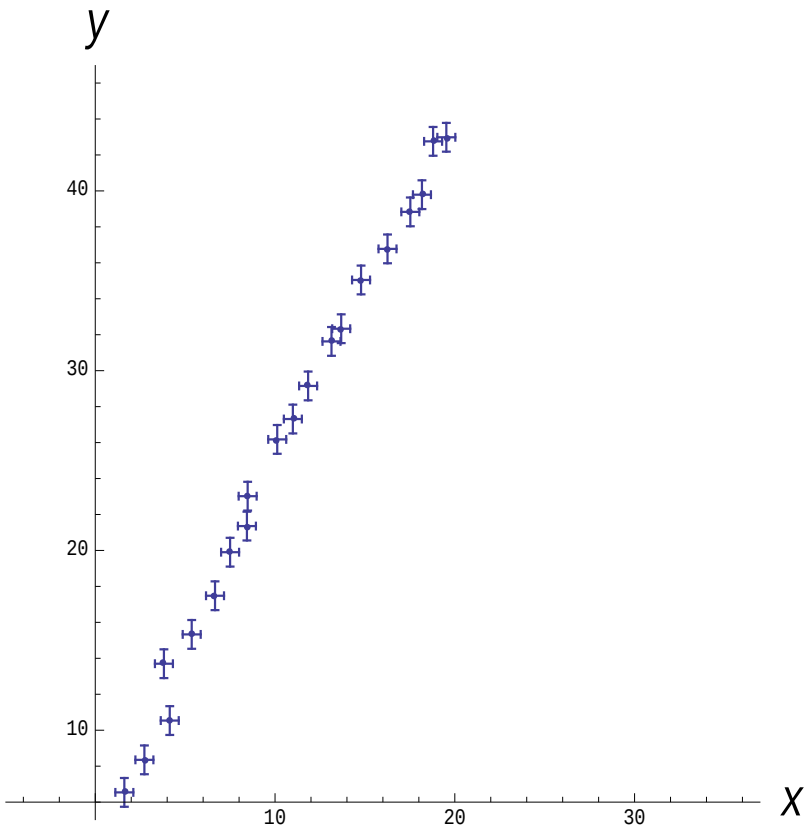

Abbildung 52: Data with errors in both variables. The goal is to find a straight line fit taking the errors in both variables into account. For this data, $\sigma_x = 0.5$ and $\sigma_y = 0.8$ are known.

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

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

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*.[8] 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. 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

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

[8] 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.

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.[9]

For the remaining prior, we use the product of gaussians

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

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. Only 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

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

Integrating out 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

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

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

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

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

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

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

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

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

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. 53.

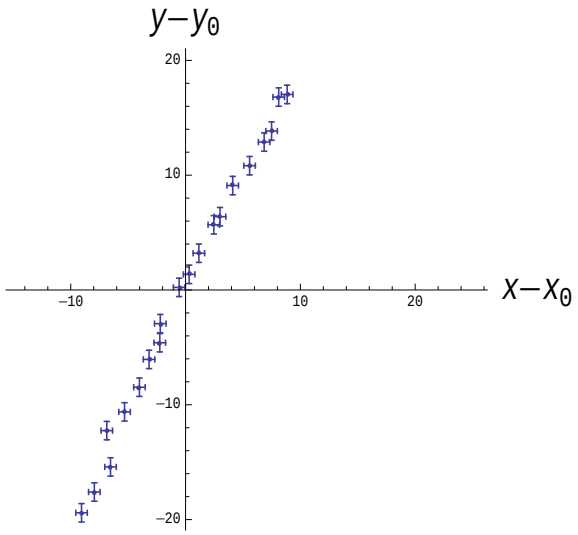

Abbildung 53: Data shifted by $x_0 = \overline{x_{\mathrm{m}i}} = 10.68$ and $y_0 = \overline{y_{\mathrm{m}i}} = 25.95$.

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

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

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

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

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. 54. We may check our result 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. 55.

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

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

It leads to

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

The minimum of the negative logarithm of this function shown in Fig. 56 is a direct way to estimate α' and its uncertainty.

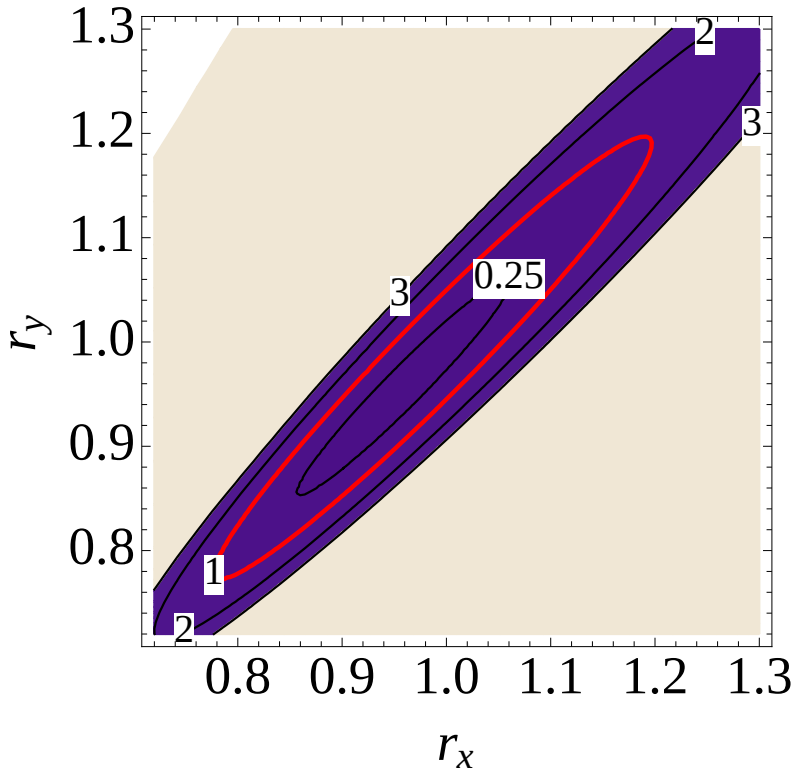

Abbildung 54: Logarithm of the posterior distribution for r_x and r_y. The minimum of this function is at $r'_x = r'_y = 0.95$. The red contour line at the value 1 may be used for estimating the standard errors in these quantities.

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

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

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

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

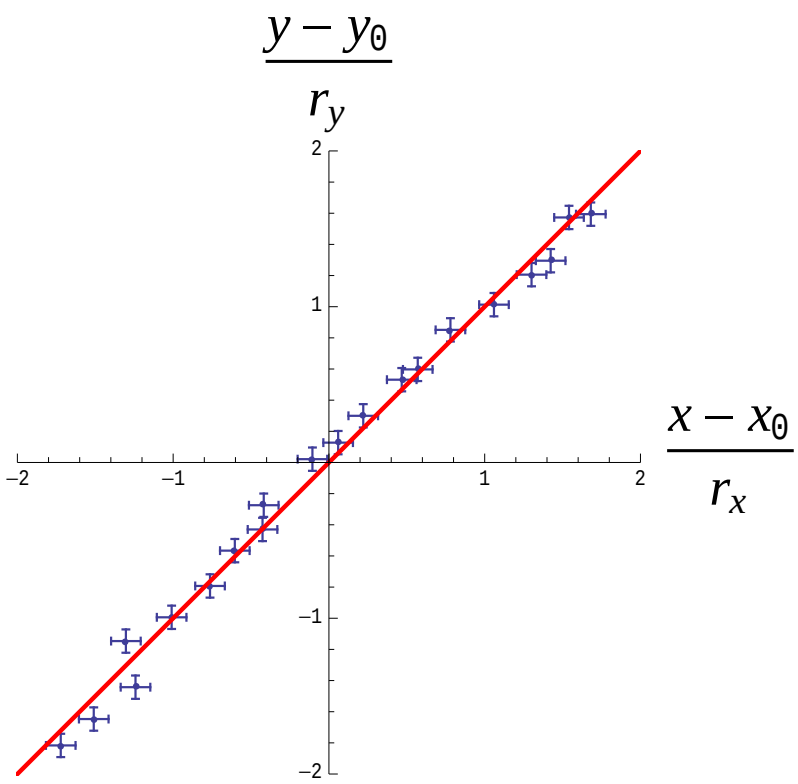

Abbildung 55: The original data shifted by (x_0, y_0) and scaled by the estimated r_x and r_y. The red line corresponds to the diagonal along which the data are expected to be scattered.

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. 57.

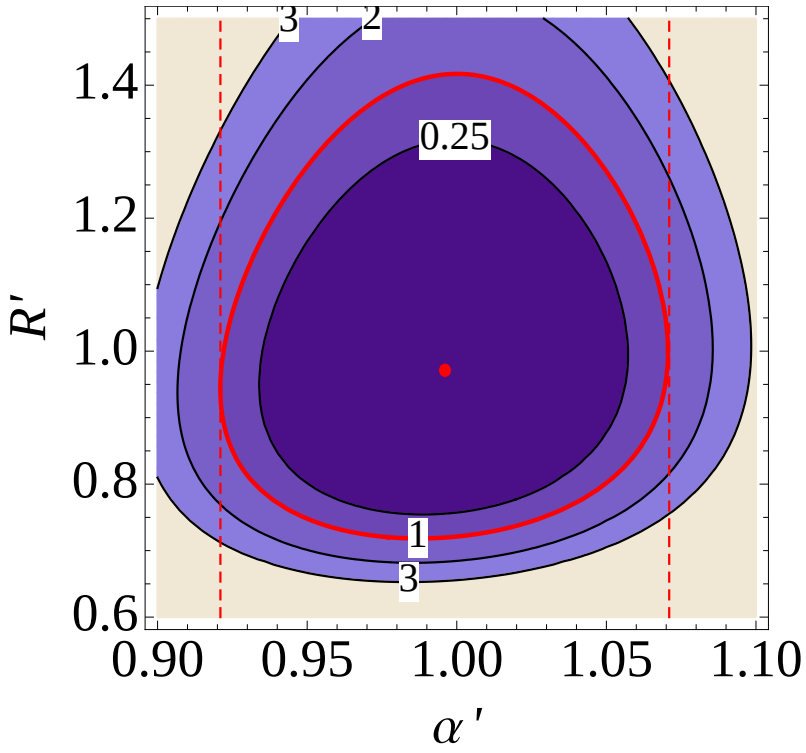

Abbildung 56: Negative logarithm of the posterior distribution for α' and R'. The minimum in this figure is found at $\alpha' = 0.996$ and $R' = 0.971$. Together with the data variance $\mathrm{Var}(x_{\mathrm{m}i}) = 30.57$ and $\mathrm{Var}(y_{\mathrm{m}i}) = 127.00$ this gives an estimate of the slope $\alpha = 2.03$. 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 $\alpha = 2.03 \pm 0.15$.

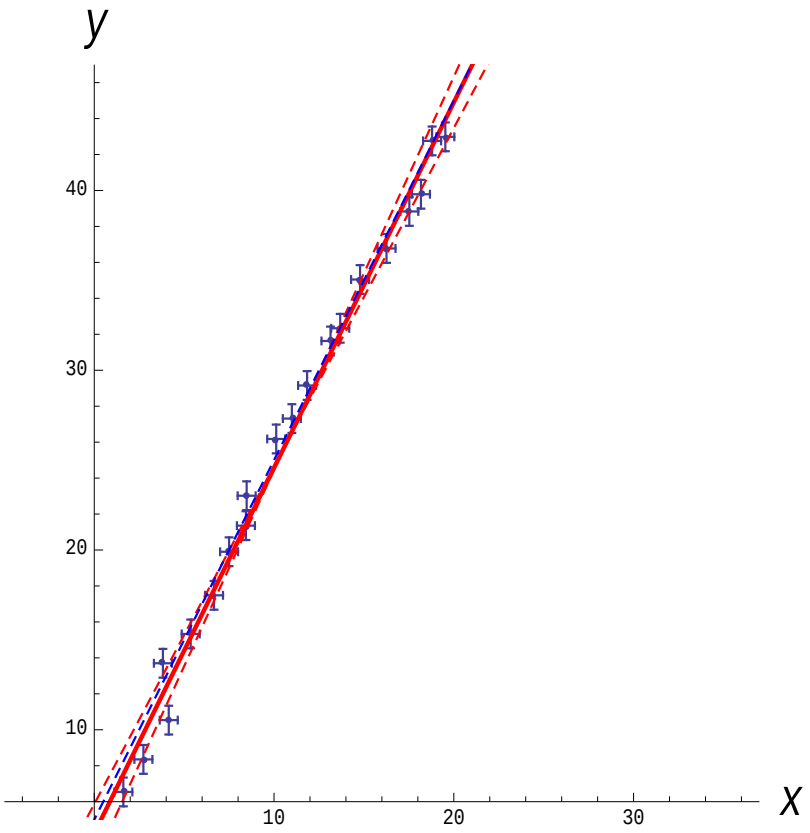

Abbildung 57: Final result of the fit to data with errors in x and y. The solid red line represents the fit with $\alpha = 2.03$ and $\beta = 4.3$. The error in β is $\Delta\beta = 1.6$. The red dashed lines have slopes $\alpha \pm \Delta\alpha = 2.03 \pm 0.15$ and $\beta = 4.3$. The blue dashed line represents the 'true' curve $y = 2x + 5$ from which the data have been generated.