Chi-square fitting

Robert Estalella

2012 June

1 Chi-square fit

Let us assume that we have \( n \) observations \( y_{i}^{\text{obs}} \) at points \( x_{i}, i = 1, \ldots, n \). The values \( x_{i} \) are supposed to be exact, while the rms error associated with each point observed \( y_{i}^{\text{obs}} \) is \( \sigma_{i} \) (and the errors are supposed to be Gaussian).

We want to fit a model to the data, \( y^{\text{mod}}(x; p_{1}, \ldots, p_{m}) \), which depends on \( x \), and on \( m \) parameters \( p_{k}, k = 1, \ldots, m \). Fitting the data means to find the set of parameters that "best fits" the observations. This "best fit" is obtained by minimizing

\[
\chi^{2} = \sum_{i=1}^{n} \left[ \frac{y_{i}^{\text{obs}} - y^{\text{mod}}(x_{i}; p_{1}, \ldots, p_{m})}{\sigma_{i}} \right]^{2}.
\]

When the minimum of \( \chi^{2} \) is found, and assuming that the model fits correctly the data (see §2), the residuals are Gaussian variables with zero mean and unit variance, and thus, \( \chi^{2} \) follows a chi-square probability distribution for \( \nu = n - m \) degrees of freedom.

2 Goodness-of-fit

The goodness of fit can be assessed by the probability of the chi-square to be higher than a given value \( \chi^{2} \) by chance, \( Q(\chi^{2} | \nu) \), where \( \nu \) is the number of degrees of freedom.

If \( \chi^{2} \gg \nu \), the probability is low (\( Q(\chi^{2} | \nu) \ll 1 \)), and the discrepancies between the model and the observations are not by chance, but have probably to be attributed to (i) a bad model; or (ii) the errors \( \sigma_{i} \) are seriously underestimated; or (iii) the errors are not Gaussian (see the discussion in Press et al. 1992).

When \( \chi^{2} \simeq \nu \), the probability \( Q(\chi^{2} | \nu) \) is high, and the model can be considered as believable. If \( \chi^{2} \ll \nu \), the model is too good to be true (or the errors are grossly overestimated).

The probability \( Q(\chi^{2} | \nu) \) can be found tabulated in most books on statistics (for instance, Wall & Jenkins 2003), or can be computed as the incomplete gamma function \( Q(a, x) \)

\[
Q(a, x) = Q \left( a = \frac{\nu}{2}, x = \frac{\chi^{2}}{2} \right) = \frac{1}{\sqrt{\nu}} \int_{x}^{\infty} e^{-t} t^{a-1} dt,
\]

which can be computed with the routine \texttt{gamq} of Numerical Recipes (Press et al. 1992).

When the number of degrees of freedom is high (for instance, \( \nu > 100 \)), the chi-square probability distribution becomes Gaussian, with mean \( \nu \) and standard deviation \( \sqrt{2\nu} \). In this case, the probability \( Q(\chi^{2} | \nu) \) is best computed as the complementary error function for the Gaussian variable of zero mean and variance \( 2 \), \( t = (\chi^{2} - \nu) / 2\sqrt{\nu} \),

\[
Q(\chi^{2} | \nu) = \frac{1}{2} \text{erfc} \left( \frac{t = \chi^{2} - \nu}{2\sqrt{\nu}} \right) = \frac{1}{\sqrt{\pi}} \int_{t}^{\infty} e^{-x^{2}} dx,
\]
which can be computed with the routines erfc or erfcc of Numerical Recipes (Press et al. 1992).

3 The case of unknown errors $\sigma_i$

If the errors of the observations, $\sigma_i$, are unknown, and we believe that the model represents well the observations (i.e. we believe that the goodness of fit is high), the discrepancies between the observations and the model for the best fit set of parameters, give us information about the error of the observations. The error that can be attributed to the measured values (the same for all) is given by

$$\sigma^2 = \frac{1}{n-m} \sum_{i=1}^{n} \left[ y_{i}^{\text{obs}} - y^{\text{mod}}(x_i; p_1, \ldots , p_m) \right]^2 .$$

4 Uncertainty in the parameter values

Once the parameters of the best fit model have been found, let us estimate the uncertainty, $\sigma(p_k)$, in the values derived for the parameters $p_k$, $k = 1, \ldots , m$.

4.1 Covariance

In some cases the covariance array can be estimated. It is the case, for instance, when the model depends linearly on the parameters,

$$y^{\text{mod}}(x_i; p_1, \ldots , p_m) = \sum_{k=1}^{m} f_k(x) p_k,$$

where $f_k(x) = \partial y^{\text{mod}} / \partial p_k$. The Jacobian array $J$ with $n$ (number of data points) rows and $m$ (number of parameters) columns is

$$J = \begin{pmatrix} f_k(x_i) / \sigma_i \\ \vdots \\ f_n(x_n) / \sigma_n \end{pmatrix} ,$$

$$J = \begin{pmatrix} f_1(x_1) / \sigma_1 & \cdots & f_m(x_1) / \sigma_1 \\ \vdots & \ddots & \vdots \\ f_1(x_n) / \sigma_n & \cdots & f_m(x_n) / \sigma_n \end{pmatrix} .$$

The array $J^T J$ is a symmetric $m \times m$ square array, given by

$$J^T J = \begin{pmatrix} \sum_{i=1}^{n} f_1(x_i) f_1(x_i) / \sigma_1^2 & \cdots & \sum_{i=1}^{n} f_1(x_i) f_m(x_i) / \sigma_1^2 \\ \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} f_1(x_i) f_m(x_i) / \sigma_n^2 & \cdots & \sum_{i=1}^{n} f_m(x_i) f_m(x_i) / \sigma_n^2 \end{pmatrix} ,$$

and its inverse is called the covariance array $C = (J^T J)^{-1}$. The uncertainties in the parameter values are given by the diagonal elements of the covariance array,

$$\sigma^2(p_k) = C_{kk}, \quad k = 1, \ldots , m .$$

When the covariance array cannot be calculated, the uncertainties $\sigma_k$ have to be estimated by means of other methods.
4.2 Monte Carlo

One approach is the Monte Carlo method. We add to the observation points $y_i^{\text{obs}}$ a Gaussian noise of zero mean and standard deviation equal to their rms error $\sigma_i$, and calculate for each trial the parameter values. The rms deviation of the parameter values obtained is the rms uncertainty $\sigma(p_k)$ of each parameter.

However, if we add a noise $\sigma_i$ to the data, the error of the data points are in fact multiplied by a factor of $\sqrt{2}$ (the variance is doubled). Thus, it seems that in this way we are overestimating the values of $\sigma(p_k)$.

4.3 Chi-square increment

A different approach is to calculate the value of $\chi^2$ for different values of the parameters, around the best fit values, for which $\chi^2$ is minimum, $\chi^2 = \chi^2_{\text{min}}$. The region of the $m$-dimensional space parameter for which $\chi^2$ does not exceed the minimum value by a certain amount $\chi^2_{\text{min}} + \Delta$, can be taken as the confidence region of the parameters. The projection of the confidence region over each axis gives the uncertainty in the parameter, $\sigma(p_k)$. What amount, $\Delta$, of the increment of $\chi^2_{\text{min}}$ is correct?

Following Avni (1976) and Wall & Jenkins (2003), the probability

$$\text{Prob} [\chi^2 - \chi^2_{\text{min}} \leq \Delta(m, \alpha)] = \alpha,$$

is that of a chi-square distribution with $m$ degrees of freedom, where $m$ is the number of parameters fitted simultaneously, and $\alpha$ is the significance level (0 < $\alpha$ < 1). The values of $\Delta(m, \alpha)$ can be found in Table 1, or calculated from the incomplete gamma function (see §2). Thus, $\Delta(m, \alpha)$ is the increment of $\chi^2$ such that if the observation is repeated a large number of times, a fraction $\alpha$ of times the values of the parameters fitted will be inside the confidence region, i.e. in the interval $p_k \pm \sigma(p_k)$. For instance, if we are fitting two parameters simultaneously, for a significance level of 0.68 (equivalent to a $\sigma$ for a Gaussian error distribution), the value of the increment in $\chi^2_{\text{min}}$ is $\Delta(2, 0.68) = 2.29$.

In short, the uncertainty in each parameter $p_k$ can be estimated as the projection of the confidence region of the $m$-dimensional parameter space over each parameter axis, $k = 1, \ldots, m$. Note that $\Delta(m, \alpha)$ depends only on the number of parameters fitted simultaneously, $m$, and not on the the goodness of fit ($\chi^2_{\text{min}}$), nor on the number of data points $n$.

The boundary of the confidence region is given by the parameter values that make

$$\chi^2 = \chi^2_{\text{min}} + \Delta(m, \alpha).$$

This condition can also be given in terms of the (weighted) rms fit residual, $\epsilon$,

$$\epsilon^2 = \sum_{i=1}^{n} \left[ (y_i^{\text{obs}} - y_i^{\text{mod}})/\sigma_i \right]^2 = \chi^2 / \sum_{i=1}^{n} 1/\sigma_i^2.$$

Assuming that the model fits well the observations, i.e. $\chi^2_{\text{min}} \simeq n - m$, we have

$$\epsilon_{\text{min}}^2 \simeq \frac{n - m}{\sum_{i=1}^{n} 1/\sigma_i^2},$$

and we obtain that the confidence region is given by the parameter values that increase the rms fit residual to

$$\epsilon \simeq \epsilon_{\text{min}} \sqrt{1 + \frac{\Delta(m, \alpha)}{n - m}}.$$
Table 1: Values of $\Delta(m, \alpha)$ for calculating the parameter uncertainties, where $m$ is the number of parameters fitted simultaneously, $\alpha$ is the significance level, given in per cent and in the equivalent number of sigmas for a Gaussian error distribution.

<table>
<thead>
<tr>
<th>$m$</th>
<th>68% (1 $\sigma$)</th>
<th>90% (1.64 $\sigma$)</th>
<th>99% (2.57 $\sigma$)</th>
<th>99.5% (2.81 $\sigma$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td>2.71</td>
<td>6.63</td>
<td>7.88</td>
</tr>
<tr>
<td>2</td>
<td>2.29</td>
<td>4.61</td>
<td>9.21</td>
<td>10.60</td>
</tr>
<tr>
<td>3</td>
<td>3.52</td>
<td>6.25</td>
<td>11.34</td>
<td>12.84</td>
</tr>
<tr>
<td>4</td>
<td>4.72</td>
<td>7.78</td>
<td>13.28</td>
<td>14.86</td>
</tr>
<tr>
<td>5</td>
<td>5.88</td>
<td>9.24</td>
<td>15.09</td>
<td>16.75</td>
</tr>
<tr>
<td>6</td>
<td>7.03</td>
<td>10.64</td>
<td>16.81</td>
<td>18.55</td>
</tr>
<tr>
<td>7</td>
<td>8.17</td>
<td>12.02</td>
<td>18.48</td>
<td>20.28</td>
</tr>
<tr>
<td>8</td>
<td>9.30</td>
<td>13.36</td>
<td>20.09</td>
<td>21.95</td>
</tr>
<tr>
<td>9</td>
<td>10.42</td>
<td>14.68</td>
<td>21.67</td>
<td>23.59</td>
</tr>
<tr>
<td>10</td>
<td>11.53</td>
<td>15.99</td>
<td>23.21</td>
<td>25.19</td>
</tr>
<tr>
<td>11</td>
<td>12.64</td>
<td>17.27</td>
<td>24.72</td>
<td>26.76</td>
</tr>
<tr>
<td>12</td>
<td>13.74</td>
<td>18.55</td>
<td>26.22</td>
<td>28.30</td>
</tr>
<tr>
<td>13</td>
<td>14.84</td>
<td>19.81</td>
<td>27.69</td>
<td>29.82</td>
</tr>
<tr>
<td>14</td>
<td>15.93</td>
<td>21.06</td>
<td>29.14</td>
<td>31.32</td>
</tr>
<tr>
<td>15</td>
<td>17.02</td>
<td>22.31</td>
<td>30.58</td>
<td>32.80</td>
</tr>
<tr>
<td>16</td>
<td>18.11</td>
<td>23.54</td>
<td>32.00</td>
<td>34.27</td>
</tr>
</tbody>
</table>

This last expression can be used even when the errors of the observations are unknown, and the rms residual is estimated as

$$
\epsilon^2 = \frac{1}{n} \sum_{i=1}^{n} (y_{i}^{\text{obs}} - y_{i}^{\text{mod}})^2.
$$

5 Numerical example

Fit of a straight line $y = px + q$, to 30 points in three different intervals, $-5 \leq x \leq 5$, $0 \leq x \leq 10$, and $10 \leq x \leq 20$, plus a Gaussian noise of mean 0 and standard deviation 1. The parameters used for calculating the data points were $p = 0.7$, $q = 3.0$. The $\chi^2$ as a function of the parameters of the fit, $p$ and $q$ is shown in Fig. 1, and the results of the fit appear in Table 2.

As can be seen in Fig 1, while for the centered interval, $-5 \leq x \leq 5$, the parameters $p$ and $q$ are independent, for the non-centered intervals, $0 \leq x \leq 10$, and $10 \leq x \leq 20$, both parameters are correlated. A best tool to fit the parameters would be a principal component analysis (Press et al. 1992). The error in $p$ does not depend on the $x$ interval, while, on the contrary, the error in $q$ depends on the $x$ interval, increasing with $\langle x \rangle$.

The 100 results obtained in the Montecarlo method (small red crosses in Fig. 1) to estimate the parameter errors are spread over the space parameters, with roughly a 70% of the points falling inside the 1-$\sigma$ (inner) contour.

The covariance and Montecarlo errors are lower than the $\chi^2$ errors. For $p$, $\sigma(p) \approx 0.06$ for the covariance and Montecarlo errors and $\sigma(p) \approx 0.09$ for the $\chi^2$ error. The errors in $q$ follow the same trend, with similar values for the covariance and Montecarlo errors, and higher values for the $\chi^2$ error. In general, the $\chi^2$ error is a $\sim 50\%$ higher than the other errors. The reason for this is clear
for the cases where \( \langle x \rangle \neq 0 \). The confidence region is not aligned with the \( p \) and \( q \) axes, and its projection onto the axes are “upper limits” of the size of the confidence region.

Table 2: Results from a fit of a straight line \( y = px + q \), to 30 points in the intervals indicated, plus a Gaussian noise of mean 0 and standard deviation 1. The covariance error was obtained from the covariance matrix, the Montecarlo error was obtained from rms of 100 trials, and the chi-square error was obtained from the 1-\( \sigma \) confidence region, for which the \( \chi^2 \) value was less than \( \chi^2_{\text{min}} + \Delta(2, 0.68) \) (see Fig. 1).

<table>
<thead>
<tr>
<th></th>
<th>Interval (-5 \leq x \leq 5)</th>
<th>Interval (0 \leq x \leq 10)</th>
<th>Interval (10 \leq x \leq 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \nu )</td>
<td>28</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>( \chi^2_{\text{min}} )</td>
<td>31.24</td>
<td>31.98</td>
<td>28.00</td>
</tr>
<tr>
<td>( Q(\chi^2_{\text{min}}</td>
<td>28) )</td>
<td>0.31</td>
<td>0.28</td>
</tr>
<tr>
<td>( p )</td>
<td>0.710</td>
<td>0.698</td>
<td>0.681</td>
</tr>
<tr>
<td>( q )</td>
<td>2.93</td>
<td>2.93</td>
<td>3.30</td>
</tr>
<tr>
<td>Cov. error</td>
<td>0.062</td>
<td>0.052</td>
<td>0.065</td>
</tr>
<tr>
<td>Mont. error</td>
<td>0.058</td>
<td>0.067</td>
<td>0.060</td>
</tr>
<tr>
<td>( \chi^2 ) error</td>
<td>0.093</td>
<td>0.091</td>
<td>0.090</td>
</tr>
<tr>
<td>( p )</td>
<td>0.28</td>
<td>0.53</td>
<td>1.38</td>
</tr>
<tr>
<td>( q )</td>
<td>0.09</td>
<td>0.37</td>
<td>0.91</td>
</tr>
</tbody>
</table>

References


5
Figure 1: Plot of $\chi^2$ as a function of the parameters of the fit, $p$ and $q$, in the fit to 30 points in the intervals $-5 \leq x \leq 5$ (top); $0 \leq x \leq 10$ (middle); and $10 \leq x \leq 20$ (bottom) (see text). The blue contours, in order of distance from the center, show the significance levels 0.68 (1 $\sigma$), 0.900, 0.990, and 0.999. The vertical and horizontal white lines show the best fit parameters obtained. The vertical and horizontal blue lines show the projection of the 1-$\sigma$ confidence region onto the $p$ and $q$ axes. The errors bars at the center cross at the best fit values, and show the errors obtained from the covariance array. The small red crosses show the values obtained for 100 runs of the Montecarlo method to find the parameter errors.