Misfit Function for Bounding Data
Bill Menke, November 22, 2020

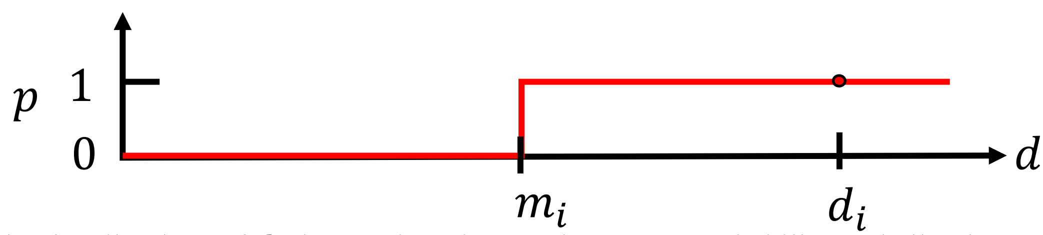
Suppose that a datum d_i is measured to accuracy σ_i and that it represents an upper bound on a model parameter m_i; that is, $d_i \geq m_i$. The issue that I consider is now to quantify the misfit function $e_i(d_i, m_i)$ so that the total error is $E = \sum_i e_i(d_i, m_i)$.

My derivation is based on the observation that the misfit can be related to the conditional probability $p(d_i|m_i)$ of the data given the model, as (Menke, 2018, Equation 9.6):

$$E = -2 \ln \prod_i p(d_i|m_i) = -2 \sum_i \ln p(d_i|m_i)$$

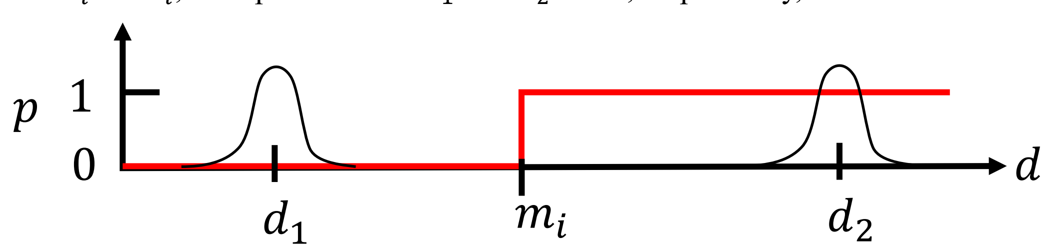
and consequently $e_i(d_i, m_i) = -2 \ln p(d_i|m_i)$. Note that in the case of the Normal p.d.f. $p(d_i|m_i) \propto \exp\{-½\sigma_i^{-2}(d_i - m_i)^2\}$ that $E = \sum_i\{\sigma_i^{-2}(d_i - m_i)^2\}$ is the usual least squared error.

In the upper bound case with noise-free data, $p(d_i|m_i) = H(d_i - m_i)$, where $H(.)$ is the Heaviside step function:

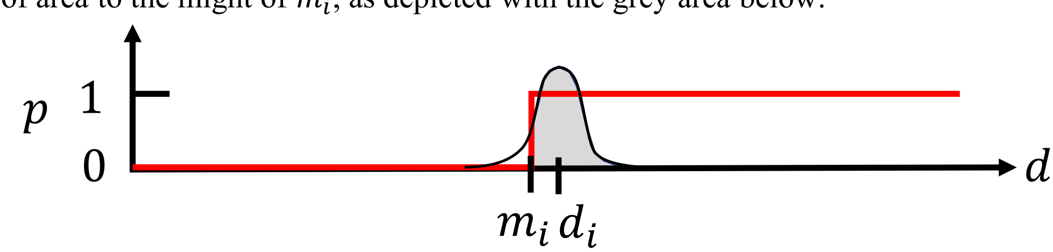


That is, all values of d_i that are less than m_i have zero probability and all values of d_i that are greater than m_i have the same non-zero probability. I note that this p.d.f. is un-normalizable, but that's not a problem here. Because of the logarithm in the definition of E, only relative probabilities affect the minimum of E. Thus I am free to define the maximum to be unity.

When d_i is Normally distributed, I still want $p(d_i|m_i) = 0$ when $d_i \ll m_i$ and $p(d_i|m_i) = 1$ when $d_i \gg m_i$, as depicted in the d_1 and d_2 cases, respectively, below.



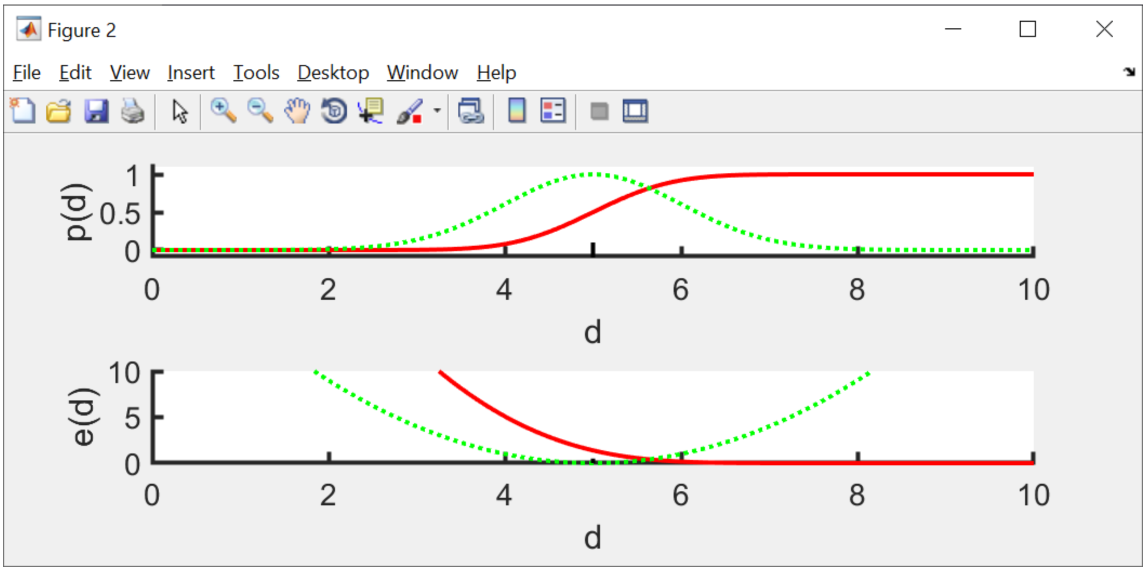
However, when $d_i \approx m_i$, it makes sense to define $p(d_i|m_i)$ to be proportional to the amount of area to the might of m_i, as depicted with the grey area below:



Consequently, we have:

$$p(d_i|m_i) = ½ + ½\,\mathrm{erf}\left(c\frac{d_i - m_i}{\sigma_i}\right) \quad \text{and} \quad e_i = -2\,\ln\left\{½ + ½\,\mathrm{erf}\left(c\frac{d_i - m_i}{\sigma_i}\right)\right\} \text{ with } c = 1$$

Note that this choice also obeys the desired limits $p \to 0$ as $(d_i - m_i) \to -\infty$ and $p \to 1$ as $(d_i - m_i) \to +\infty$. These functions are plotted in red in the graph, below (with the green curve showing the result for a Normal $p(d_i|m_i)$ of the same variance).



The lower bound corresponds to $c = (-1)$ in the above equation. In the following example, a grid search is used to determine the intercept a and slope b of a straight line, subject to upper bound (green), lower bound (red) and point data (cyan). The true values are $(a, b) = (1,2)$, the true line is shown in black, the estimate values are (1.47,1.98) and the best fit line is shown in blue. The example uses uniform variance σ_i=1.