# Details

beta fitter

## 1 Beta distribution:

The program uses the beta distribution given by the Scipy library with 4 parameters: **a**, **b**, $\mathbf{x_{min}}$ and $\mathbf{x_{max}}$ (or respectively loc and scale). Those parameters define the beta distribution as:

$$\beta(y, a, b, x_{max}) = \frac{\Gamma(a+b)y^{a-1}(1-y)^{b-1}}{\Gamma(a)\Gamma(b)x_{max}} \tag{1}$$

where **y** is defined as the standardized variable:

$$y = \frac{x - x_{min}}{x_{max} - xmin} \in [0, 1] \tag{2}$$

and the gamma function is:

$$\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}dt, \quad z \in \mathbb{R}_+ \tag{3}$$

The cumulative density function of the beta distribution (Also known as regularized incomplete beta function) takes the form:

$$C_y(a, b) = \frac{B(y, a, b)}{B(1, a, b)} \tag{4}$$

With $B(y, a, b)$ the Beta function defined as:

$$B(y, a, b) = \int_0^y t^{a-1}(1-t)^{b-1}dt \tag{5}$$

## 2 Fitting algorithm:

**Fit:** The fit is done on the cumulative distribution function of the beta distribution. The fitting algorithm is based on minimizing the Mean Square Error (MSE) between the data and the fit with the *minimize* method from Scipy, which uses in this case the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm. In particular, it uses a limited-memory version of this algorithm under boxes constraint (L-BFGS-B), where boxes constraints are linear constraints put on the parameters of the fit. In the case of the application, the constraints are set on the parameters which handle the support of the cumulative distribution function i.e. $x_{min}$ and $x_{max}$.

**Constraints:** The application handle the data such that if the boundaries are given in the dataset the algorithm doesn't optimize over them: if the data contains the point ($I_1$, $V_1 = 0$) or ($I_N$, $V_N = 1$) it fixes respectively $x_{min} = I_1$ or $x_{max} = I_N$. Otherwise the optimization is made with the constraints $0 < x_{min} \leq I_1$ or $I_1 - 0.1(I_N - I_1) < x_{min} \leq I_1$ in the case $0 < I_1 - 0.1(I_N - I_1)$, and $I_N < x_{max} < I_N + 0.1(I_N - I_1)$. The "margins", fixed here at 10% of the length of interval of observed intensities avoids $x_{min}$ and $x_{max}$ to diverge or get non-coherent values when optimizing on them. The MSE is also slightly modified by adding penalty terms on the boundaries: $0.1C_0(a, b)^2$ and $0.1(1 - C_1(a, b))^2$ to enforce the cumulative to go to 0 on the left and 1 on the right.

**Interval of confidence (IC):** For the interval of confidence, the distance between the points and the fit is considered to follow a Gaussian distribution. The parameters of their distribution are estimated, this time using the *stats.norm.fit* method from scipy, and the 0.99 quantile of the distribution is used as superior and inferior limits around the beta fit. Naturally the intervals of confidence are bounded to 0 and 1, since we know that the cumulative distribution function has to be inside this interval. It is good to remark that the initial assumption for the IC is not true, and that determining the upper and lower error by up-lifting and down-lifting the fit is a naive method for determining the continuous IC.