Visual Averaging Library Code

Michael Birch
McMaster University

Overview

- Basic features
  - Treatment of asymmetric uncertainties
  - Outlier detection
  - Eight averaging procedures
  - Multiple data set handling
- Data visualization
  - Gaussian View
  - Point view

Outlier Detection

- Chauvenet's Criterion
  - Does not consider uncertainties of individual data points
  - Sets arbitrary cut-off for deviation from the unweighted average
  - Single outlier identified, iterated for multiple outlier detection; this is very dangerous!
- Peirce's Criterion
  - Again does not consider uncertainties
  - Based on principle that a subset of points should be outliers if the likelihood of the entire data set is less than the likelihood of the data set with the subset removed multiplied by the probability of the existence of that number of outliers

Outlier Detection

- Birch's Criterion
  - Considers uncertainties
  - Identifies outliers with respect to a particular recommended result if the difference is not consistent with 0 within a specified confidence level

Averaging Procedures

- Unweighted Average
- Weighted Average
- Limitation of Statistical Weighted (LWM)
- Normalized Residuals Method (NRM)
- Rajeval Technique (RT)
- Method of Best Representation (MBR)
- Bootstrap, with uncertainties considered
- Mandel-Paule Method

Data Visualization

- Gaussian View
  - Each measurement and associated uncertainty is used to define a Gaussian (Normal) distribution for the data point
  - These distributions may be summed and re-normalized to obtain a "mean probability density function" to represent the entire data set
  - Results of different averaging methods also viewed at Gaussian distributions may be compared with this representation

Data Visualization

- Point View
  - Each measurement plotted as a point with error bars indicating the uncertainty
  - The different averaging results may be compared with the data by adding them to this plot as a line with shaded band indicating the uncertainty

Chauvenet's Criterion (in manual of practical astronomy)

- William Chauvenet decided (circa 1860) an "outlier" should be defined as a value in a set of n measurements for which the deviation from the mean, $d_i = |x_i - \bar{x}|$, would be observed with probability less than $1/2n$ assuming the data are distributed according to a normal distribution with the sample mean, $\bar{x}$, (unweighted average) and variance, s^2, given by the unbiased sample variance (a quantity defined in any statistics text). Iterative approach with one outlier picked up at a time

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- Note that the uncertainties of the individual data points are not taken into account
- A formula for the criterion is thus obtained by the following calculation

$$\Pr(X \geq \bar{x} + d_i) + \Pr(X \leq \bar{x} - d_i) < \frac{1}{2n}$$

$$\int_{\bar{x}+d}^{\infty} \mathcal{N}(x;\bar{x},s)dx + \int_{-\infty}^{\bar{x}-d} \mathcal{N}(x;\bar{x},s)dx < \frac{1}{2n}$$

$$1 - \operatorname{erf}\left(\frac{d_i}{\sqrt{2}s}\right) < \frac{1}{2n}$$

$$n \cdot \operatorname{erfc}\left(\frac{d_i}{\sqrt{2}s}\right) < \frac{1}{2}$$

where $\operatorname{erf}(x)$ is the "error function" defined by

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$$

and $\operatorname{erfc}(x)$ is the complimentary error function defined by $\operatorname{erfc}(x) = 1 - \operatorname{erf}(x)$

Peirce's Criterion

- Benjamin Peirce developed a criterion for finding outliers a few years before Chauvenet and his work is more mathematically rigorous
- He assumes the data are distributed according to the same normal distribution as Chauvenet, however the principle used to identify outliers is very different
- A subset of m points are called outliers if

  (likelihood of the complete data set) < (likelihood of the remainder of the data set)(Probability of the existence of m outliers)
- The principle includes the identification of more than one outlier hence the procedure for identifying outliers need not be iterated as with Chauvenet's criterion
- It is difficult to obtain an exact, closed form solution to the inequality above using the appropriate likelihood functions; however an iterative procedure can be used to find the maximum deviation from the mean, above which the measurements can be considered outliers by the above principle

Peirce's Criterion

- After working with the mathematical formulation for Peirce's principle the following four equations are derived to obtain the ratio of the maximum deviation from the unweighted mean, d_{max}, to the square root of the sample variance, s, as defined for Chauvenet's Criterion: $r_{max}=d_{max}/s$.
  - Suppose in a set of n measurements m are suspected as outliers

$$Q^n = \frac{m^m (n-m)^{n-m}}{n^n} \qquad (1)$$

$$\lambda^{n-m} R^n = Q^n \qquad (2)$$

$$r_{\max}^2 = \lambda^2 + \frac{n-1}{m}(1-\lambda^2) \qquad (3)$$

$$R = e^{\frac{1}{2}\left(r_{\max}^2 - 1\right)} \operatorname{erfc}\left(\frac{r_{\max}}{\sqrt{2}}\right) \qquad (4)$$

  - These lend themselves to the iterative procedure to find r_{max}
    - 1. Calculate Q using equation (1)
    - 2. Begin with an approximate value for R
    - 3. Use Q and R to calculate λ by equation (2)
    - 4. Use λ to calculate r_{max} using equation (3)
    - 5. Use r_{max} to refine the estimate on R using equation (4)
    - Repeat steps 3-5 until R converges to one value, the r_{max} which gives that R is the required maximum ratio

Peirce's Criterion

- To apply Peirce's criterion:
  - First assume one point is an outlier ($m=1$), then check if that is true by checking if any points exceed the maximum deviation from the unweighted mean calculated as on the previous slide
  - If there are any outliers then assume there is one more (for example if 3 points exceed the maximum deviation then try the calculation with $m=4$) and repeat the calculation until no more outliers are identified
- Note that even though Peirce's criterion in more rigorous than Chauvenet's and does not arbitrarily choose a probability which indicates outliers, this formulation still does not include the uncertainties of the respective data points as they are not included in the likelihood functions

Method to including uncertainties by M. Birch

- It is proposed here that a criterion for identifying outliers which takes into account the uncertainties on each data point may be defined as follows:
  - A measurement $x_i \pm \sigma_i$ is outlier with respect to a supposed mean $\mu \pm \sigma_\mu$ if the difference

  $d = x_i - \mu$ is "inconsistent with zero" at a given confidence level, α.
- It can then be proven that the random variable

  $D = X_i - M$ will be normally distributed about d with variance

  $\sigma_d^2 = \sigma_i^2 + \sigma_\mu^2$ where X_i and M are normally distributed random variables with their respective peak values at the measurement x_i and supposed mean μ
- We can say D is inconsistent with zero at a confidence level α if $\Pr(0<D<2d) > \alpha$ when $d > 0$ or $\Pr(2d<D<0) > \alpha$ if $d < 0$, since these intervals correspond to the set of values more likely to occur than zero.
- This results in the formula

$$\operatorname{erf}\left(\frac{|x_i - \mu|}{\sqrt{2\left(\sigma_i^2 + \sigma_\mu^2\right)}} \right) > \alpha$$