CEWE manual

A.C.P. Oude Nijhuis

September 22, 2016

Abstract

This is the manual of the Comparison of Everything with Everything (CEWE) tool, which is another massive data analysis tool.

1 Introduction

In this document the Correlate Everything With Everything (CEWE) tool is described, which is another massive data analysis tool in the form of a *Python* module, exploiting the concept of addition of summary statistics. The purpose of CEWE is to calculate statistics, histograms and correlations for non-overlapping subdatasets and store those results in self-descriptive hierarchical data 5 (HDF5) format files. The HDF5 is a data model, library, and file format for storing and managing data, that is for example used in satellite data Brennan *et al.* (2013). The CEWE tool provides the possibility for adding up statistics from two CEWE HDF5 files, giving a new CEWE HDF5 file containing statistics, histograms and correlations for the combined dataset. This feature makes the CEWE tool memory efficient and adaptable, which is especially useful for a massive dataset with numerous variables stored in large files in an efficient manner. The CEWE tool is efficient for processing large data files, e.g. radar data files, in a consecutive manner, i.e. without having to access all the files at once or several times. The *Python* module open source code is publicly and freely available on *GitHub* at `http://github.com/albertoudenijhuis/cewe/`. The code includes libraries for statistical operations, plotting and file operations.

In the next section some calculations that are required for the massive data analysis tool, such as the addition of summary statistics, are explained. In the consequent section it is explained how to start up a project, starting from a minimal worked example that is provided in the package.

2 Calculations

2.1 Addition of summary statistics

The addition of summary statistics is an exact solution to the problem that the calculation of statistics for massive data sets is either infeasible or impractical. We refer to the 'addition of a statistic' as a formula for the calculation of a statistic of two combined datasets with independent samples, given only the statistic of two datasets. Statistics that are used in the CEWE tool are raw moments μ_{xk} and the mixed moment ν_{xy}:

$$\mu_{xk} = \sum_i \frac{x_i^k}{n} \quad , \quad \nu_{xy} = \sum_i \frac{x_i y_i}{n}, \tag{1}$$

where x_i and y_i are the samples, i is indicating the sample number and n is the number of samples. The raw moments have the property that for two independent datasets A and B, the statistic can be calculated by only using the statistics of the two datasets:

$$\mu_{xk,AB} = \frac{n_A \mu_{xk,A} + n_B \mu_{xk,B}}{n_A + n_B}. \tag{2}$$

This property also applies to the mixed moment ν_{xy}. Given the mixed moment and the first two raw moments ($k = 1, 2$) the mean, the variance, the covariance and correlation coefficient are calculated (appendix). The mixed moment and first two moments are also used to calculated linear regression coefficients (appendix). With the addition of higher order raw moments ($k = 3, 4$) statistics such as the skewness and the kurtosis are also calculated.

2.2 Circular statistics

Circular variables arise in various ways, e.g. compass or wind direction measurements. It is tempting to apply linear statistics to circular measurements. In such a case the circular domain has to be cut with the drawback that the statistics depend strongly on the point where the circle is cut. Existing circular statistics that are cutting the domain, e.g. expressed by usage of the modulo operator, are problematic because for them it is not possible to formulate a rule for the addition of a statistic. Therefore the embedding approach is used, which is a solution from Mardia Mardia & Jupp (1999), where linear statistics are applied to a vector $(\cos\theta, \sin\theta)$ for a circular variable θ (details in the appendix). The advantage of this approach is that the calculations are linear and the domain is not cut. Also note that the embedding approach can be applied to any periodic variable by rescaling the domain length to 2π. For more details on directional statistics we refer to book on directional statistics from Mardia Mardia & Jupp (1999).

2.3 2D histograms

A histogram is a tally of the data for intervals. When the intervals and domain are properly chosen, the histogram resembles the probability density function of the data which can be used for analysis of the data. In CEWE equidistant intervals are used which suffices when the upper limit, lower limit and number of intervals are chosen properly. Histograms of two data sets can be added when the chosen intervals are kept the same. When two variables are compared, a 2D histogram is calculated consisting of 2D grid boxes which are made from the 1D equidistant intervals. Again the 2D histograms can be added for two datasets when the 2D grid boxes are kept the same. Given a normalized 2D histogram, scatter density plots can be made to inspect data correlations. Such scatter density plots can be useful to visually inspect if two variables have a linear or a more complex relation.

2.4 Calculation of other statistics

With some effort the CEWE tool can be extended to other statistics in two different manners. The first manner is to formulate an addition rule for the statistic, as was shown for the raw moments in Eq. 2. The second manner is the approximation of the original samples by the 2D histogram center values. The first manner is an exact solution and therefore attractive but it requires extra analysis effort. There are also cases that the first manner is not possible at all, for example when the modulo operator is encountered, e.g. in some circular statistics. The second manner requires a validation of the sample approximation, which can be done by visual

inspection of the 2D histogram. The second manner is used to plot distributions and has the potential of e.g. being used for the calculation of median and percentiles.

3 How to start up a CEWE project

CEWE can be downloaded from the GitHub website or downloaded via the commandline as:

```
git clone https://github.com/albertoudenijhuis/cewe
```

An example is provided in the CEWE package. To start a new project we suggest to copy this example as a starting point. The example consists of two files:

```
cewe_example1.py
cewe_example1_dataset_attributes.py
```

The first file is the main file where data samples are collected and statistical calculations are executed. The second file is the file where the attributes for the variables are set. In the example we can see:

```
dataset_attributes = {}

dataset_attributes['A'] = {
        'circular': False,
        'llim': -3.,
        'ulim': 3.,
        'plotname': 'A',
        'units': '-',
        'scalefactor': 1.,
        }
dataset_attributes['B'] = {
...
```

The attributes are stored in *Python* dictionaries. In the example this is done for the variables with shortnames "A" until "D". For each variables it has to be specified whether it is a circular variable or not. An example of a circular variable is the wind direction. The upper limit and lower limit are relevant for the 1D and 2D histograms. For circular calculations a period is used of the difference between the limits. The last three attributes are relevant for plotting which are the plotname, the units and the scalefactor. In a new project the shortnames can be replaced and additional variables can be added.

In the main file the dataset dictionary is made that contains all input for the CEWE calculations.

```
dataset = {}
dataset.update(dataset_attributes)
dataset['A']['samples'] = np.random.normal(size=nsamples)
...
```

The attributes are taken over from the attributes dictionary and consequently the samples are stored in the dictionary. Once all the samples are set, the CEWE HDF5 statistics file can be made via:

```
cewe.dataset2ceweh5file(dataset, 'cewe_example1_dataset1.h5')
```

Given two CEWE statistics files, representing two data sets, e.g. two measurement days, their statistics can be combined via:

```
cewe.combine_ceweh5files('cewe_example1_dataset1.h5',
'cewe_example1_dataset2.h5', 'cewe_example1_dataset1_and_dataset2.h5')
```

A scatter density plot for variable "A" and "b" can be made via:

```
cewe.scatter_density_plot('cewe_example1_dataset1_and_dataset2.h5', 'A', 'B')
```

More options can be accomblisehd by using the dictionary that holds all CEWE statistics and can be obtained via:

```
cewe_dct = read_cewe_hdf5_file(cewe_example1_dataset1_and_dataset2.h5)
```

4 Conclusion

Another massive data analysis tool is now available for free use. Some features of this massive data analysis tool are that (1) it can handle circular variables; (2) it is implemented in *Python* and; (3) it produces self-descriptive hierarchical data 5 (HDF5) format files containing statistics that can easily be recombined.

References

Brennan, J, Lee, HJ, Yang, M, Folk, M, & Pourmal, E. 2013. Working with NASA's HDF and HDF-EOS earth science data formats. *Earth Obs. Newsl.*, **25**(April 2013).

Mardia, Kanti V., & Jupp, Peter E. 1999. *Directional Statistics*. 2nd editio edn. London: Academic Press.

Rice, John A. 1995. *Mathematical Statistics and Data Analysis, Volume 1*.

Taylor, J R. 1997. *An introduction to error analysis: The study of uncertainties in physical measurements, second edition*. Sausalito, California: University Science Books.

A Calculations

A.1 Calculation of mean, variance and covariance

Given the set of variables

$$(n, \mu_{x1}, \mu_{x2}, \mu_{y1}, \nu_{xy}), \tag{3}$$

that were calculated from the samples x_i and y_i, the mean $\mathrm{Mean}(x)$ and the sample variance $\mathrm{Var}(x)$ are calculated as:

$$\mathrm{Mean}(x) = \mu_{x1} \quad , \quad \mathrm{Var}(x) = \frac{n}{n-1}\left[\mu_{x2} - \mu_{x1}^2\right], \tag{4}$$

and the sample covariance $\mathrm{Cov}(x, y)$ as:

$$\mathrm{Cov}(x, y) = \frac{n}{n-1}\left[\nu_{xy} - \mu_{x1}\mu_{y1}\right]. \tag{5}$$

The correlation coefficient $r_{x1,x2}$ can be calculated given the covariance and variance as:

$$r_{xy} = \mathrm{Cov}(x, y)/\sqrt{\mathrm{Var}(x)\mathrm{Var}(y)}. \tag{6}$$

A.2 Calculation of skewness and kurtosis

Given the first three raw moments, the skewness can be calculated asRice (1995):

$$\left\langle \left(\frac{X-\mu}{\sigma}\right)^3 \right\rangle = \frac{\mu_{x3} - 3\mu_{x1}\mu_{x2} + 2\mu_{x1}^3}{\left(\mu_{x2} - \mu_{x1}^2\right)^{3/2}}, \tag{7}$$

and given the first four raw moments, the kurtosis can be calculated asRice (1995):

$$\left\langle \left(\frac{X-\mu}{\sigma}\right)^4 \right\rangle = \frac{\mu_{x4} - 4\mu_{x1}\mu_{x3} + 6\mu_{x1}^2\mu_{x2} - 3\mu_{x1}^4}{\left(\mu_{x2} - \mu_{x1}^2\right)^2}. \tag{8}$$

A.3 Linear regression

The solution to a linear least squares fit of the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}, \tag{9}$$

where $x_1, \cdots, x_{p-1}, y$ are the variables and β are the fitted parameters, can be written in matrix notation asRice (1995):

$$\hat{\beta} = (X^T X)^{-1} X^T Y, \tag{10}$$

where

$$\hat{\beta} = \begin{pmatrix} \beta_1 \\ \cdots \\ \beta_n \end{pmatrix} \quad , \quad X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{p-1,1} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & \cdots & x_{p-1,n} \end{pmatrix} \tag{11}$$

and

$$Y = \begin{pmatrix} y_1 \\ \cdots \\ y_n \end{pmatrix}, \tag{12}$$

where the subscript i indicates the sample number for x_{ij} and y_i and j indicates the variable number in x_{ij}. We can express two terms to calculate $\hat{\beta}$ in Eq. 10 as:

$$X^T X = \begin{pmatrix} n & \sum_i x_{1i} & \cdots & \sum_i x_{ni} \\ \sum_i x_{i1} & \sum_i x_{1i}x_{1i} & \cdots & \sum_i x_{1i}x_{ni} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_i x_{in} & \sum_i x_{ni}x_{11} & \cdots & \sum_i x_{ni}x_{ni} \end{pmatrix} \tag{13}$$

and

$$X^T Y = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i1}y_i \\ \cdots \\ \sum_i x_{in}y_i \end{pmatrix}. \tag{14}$$

These terms can be calculated given the set of variables that are available in the CEWE tool. The solution for one variable ($p = 1$) can also be written more conveniently asTaylor (1997):

$$\beta_1 = \mathrm{Cov}(x_1, y)/\mathrm{Var}(x_1), \tag{15}$$

$$\beta_0 = \mathrm{Mean}(y) - \beta_1 \mathrm{Mean}(x_1), \tag{16}$$

and also for two variables ($p = 2$) the solution can be written as:

$$\beta_1 = \frac{\mathrm{cov}(x_1, x_2)\mathrm{cov}(x_1, y) - \mathrm{cov}(x_1, y)\mathrm{var}(x_2)}{\mathrm{cov}^2(x_1, x_2) - \mathrm{var}(x_1)\mathrm{var}(x_2)}, \tag{17}$$

$$\beta_2 = \frac{\mathrm{cov}(x_1, x_2)\mathrm{cov}(x_1, y) - \mathrm{cov}(x_2, y)\mathrm{var}(x_1)}{\mathrm{cov}^2(x_1, x_2) - \mathrm{var}(x_1)\mathrm{var}(x_2)}, \tag{18}$$

$$\beta_0 = \mathrm{Mean}(y) - \mathrm{Mean}(x_1)\beta_1 - \mathrm{Mean}(x_2)\beta_2. \tag{19}$$

A.4 Circular variables

For a circular variable θ, two new variables are considered the cosinus c_θ and the sinus s_θ, i.e. the Cartesian coordinates, which are calculated for each sample θ_i as:

$$c_{\theta i} = \cos\theta_i \quad , \quad s_{\theta i} = \sin\theta_i. \tag{20}$$

The variables c_θ and s_θ are linear and therefore standard statistics is applied on them in the CEWE tool. Then the statistics of c_θ and s_θ are used to calculate the statistics of the circular variable. The average values of the Cartesian coordinates are:

$$\overline{c_\theta} = \sum_i \frac{\cos\theta_i}{n} \quad , \quad \overline{s_\theta} = \sum_i \frac{\sin\theta_i}{n}, \tag{21}$$

which leads to the definition of the mean angle Mardia & Jupp (1999) :

$$\overline{\theta} = \begin{cases} \tan^{-1}(\overline{s_\theta}/\overline{c_\theta}) & \text{if } \overline{c_\theta} \geq 0, \\ \tan^{-1}(\overline{s_\theta}/\overline{c_\theta}) + \pi & \text{if } \overline{c_\theta} < 0. \end{cases} \tag{22}$$

A measure for the circular variance is given by Mardia & Jupp (1999):

$$\mathrm{Var}(\theta) = -2\log \overline{r_\theta} \text{ with } \overline{r_\theta} = \sqrt{\overline{c_\theta}^2 + \overline{s_\theta}^2}. \tag{23}$$

A measure for the dependence between a linear variable x and a circular variable θ is the sample multiple correlation coefficient $r_{x\theta}$ of x and (c_θ, s_θ), which can be calculated given ordinary correlation coefficients Mardia & Jupp (1999):

$$r^2_{x\theta} = \frac{r^2_{xc} + r^2_{xs} - 2r_{xc}r_{xs}r_{cs}}{1 - r^2_{cs}}, \tag{24}$$

where $r_{xc} = \mathrm{corr}(x, c_\theta)$, etc.. A measure for the dependence between two circular variables θ and ϕ can be based on the sum of squared canonical correlation coefficients of (c_θ, s_θ) and (c_ϕ, s_ϕ) and is given by Mardia & Jupp (1999):

$$\begin{aligned} r^2_{\theta\phi} = \big[&(r^2_{cc} + r^2_{cs} + r^2_{sc} + r^2_{ss}) + 2(r_{cc}r_{ss} + r_{cs}r_{sc})r_1r_2 \\ &-2(r_{cc}r_{cs} + r_{sc}r_{ss})r_2 - 2(r_{cc}r_{sc} + r_{cs}r_{ss})r_1]/ \\ &\big[(1 - r_1^2)(1 - r_2^2)\big], \end{aligned} \tag{25}$$

where $r_{cc} = \mathrm{corr}(c_\theta, c_\phi)$, etc., $r_1 = \mathrm{corr}(c_\theta, s_\theta)$ and $r_2 = \mathrm{corr}(c_\phi, s_\phi)$ are the ordinary sample correlation coefficients. The interpretation of $r_{x\theta}$ and $r_{\theta\phi}$ is different than for the normal correlation coefficient r_{xy}, for which we refer to Mardia Mardia & Jupp (1999). For a best fit between a circular variable θ and a linear variable y, we continue to use the linear variables, the cosinus c_θ and sinus s_θ. Therefore the problem is equivalent to finding the solution of a least squares fit with the model $y = \beta_0 + \beta_1 x_1 + \beta_1 x_2$, where x_1 equals c_θ and x_2 equals s_θ. This solution can also be used for a least squares fit of a circular variable θ against a circular variable ϕ. For such a case c_θ is fitted as function of c_ϕ and s_ϕ and for s_θ the same. The two fitted functions are then used to obtain the angle θ via the $\tan^{-1}$ function (see Eq. 22). Note that the resulting fit function is then non-linear.