pyHomogeneity: A python package for homogeneity test of time series data

10 December 2020

Summary

The homogeneity test is a statistical test method, that checks if two (or more) datasets came from the same distribution or not. In a time series, the homogeneity test is applied to detect one (or more) change/breakpoint in the series. This breakpoint occurs where the data set changes its distribution. This makes homogeneity test an important test in statistical analysis. There are several tests available to check homogeneity. These tests can be performed using several commercial software packages and different programming languages. However, a single package to perform most of the widely used homogeneity tests will save time and bring diversity in analysis.

Python is one of the widely used tools used by data scientist. A large number of data analysis and research tools are also developed using Python. But, till now there is no Python package available for the homogeneity test. `pyHomogeneity` package fills this gap. It is a pure Python implementation for the homogeneity test. `pyHomogeneity` can perform six homogeneity tests which are widely used in time series analysis. Available tests in `pyHomogeneity` package are briefly discuss in below:

Pettitt Test: In 1979, [@pettitt1979non] proposed a change-point detection test based on Mann-Whitney two-sample test. For continues dataset, Pettitt statistics U(k) can be can be calculated using follows:

$$U(k) = 2\sum_{i=1}^{n} r_i - k(n+1)$$

Where, r1, r2, r3, ……., rn are the ranks of the n observations x1, x2, x3, ……, xn in the complete sample of n observations. The maximum absolute value of U(k) is refer to the probable change point at k-th data. The approximate probability for a two-sided test is given by

$$p = 2 \exp\left(\frac{-6 * (\max(|U(k)|))^2}{n^3 + n^2}\right)$$

Where, the approximate probability is good for p 0.5 [@pettitt1979non]. The probability or critical values for the test-statistic also can estimated by using Monte Carlo simulation.

Standard Normal Homogeneity Test (SNHT): Standard Normal Homogeneity Test (SNHT) is based on the Ratio Test Method [@alexandersson1986homogeneity]. This method is best suitable to detect inhomogeneity near the beginning and end of the series [@Mahmud2015homo]. The test statistic T(k) is calculated by comparing the mean of the first k data of the record with the last n-k data as follows:

$$T(k) = k\bar{z}_1^2 + (n+k)\bar{z}_2^2$$

Where,

$$\bar{z}_1 = \frac{1}{k}\sum_{i=1}^{k}\frac{x_i - \bar{x}}{S} \bar{z}_2 = \frac{1}{n-k}\sum_{i=k+1}^{n}\frac{x_i - \bar{x}}{S} S = \text{sample standard deviation} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} n = \text{number o}$$

The test statistic T(k) reaches it's maximum value when a break point is detected at the data point K. The test statistic T0 is defined as:

$$T_0 = max(T(k))$$

The null hypothesis will be rejected if T0 is above a certain level, which is estimated by using Monte Carlo simulation.

Buishand's test: [@buishand1982some] proposed a homogeneity test method based on adjusted partial sums. The test statistics are given below:

$$S(k) = \sum_{i=1}^{k}\frac{x_i - \bar{x}}{\sigma}$$

Where,

$$\sigma = \text{standard deviation} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

The maximum absulate value of S(k) is referred to as the probable change point at k-th data. [@buishand1982some] proposed four way to check sensitivity of this homogeneity test. Those are

Q test: In this method, Calculate Q as follow and estimate Critical values for the test-statistic using given table by [@buishand1982some] or using Monte Carlo simulation.

$$Q = \frac{\max(S(k))}{\sqrt{n}}$$

Range test: In this method, Calculate the range R using the equation below and estimate Critical values for the test-statistic using given table by [@buishand1982some] or using Monte Carlo simulation

$$R = \frac{\max(S(k)) - \min(S(k))}{\sqrt{n}}$$

Likelihood Ratio test: the test statistics V(k) is calculated as following and estimate Critical values for the test-statistic using Monte Carlo simulation

$$V = \max\left(\frac{|S(k)|}{\sqrt{k(n-k)}}\right)$$

U Test: According to [@buishand1984tests], U statistics is a robust test and good to detect change point in middle of series. The U statistics is calculated as follow and estimate Critical values for the test-statistic using the given table by [@buishand1982some] or using Monte Carlo simulation

$$U = \frac{1}{n(n+1)} \sum_{i=1}^{n-1} S(k)^2$$

Example

A quick example of pyHomogeneity usage is given below.

```
import numpy as np
import pyhomogeneity as hg

# Data generation for analysis
data = np.random.rand(360,1)

result = hg.pettitt_test(data)
print(result)
```

Output are like this:

```
Pettitt_Test(h=False, cp=89, p=0.1428, U=3811.0, avg=mean(mu1=0.5487521427805625, mu2=0.4688
```

Whereas, the output is a named tuple, so user can call by name for specific result:

```
print(result.cp)
print(result.avg.mu1)
```

or, user can directly unpack the results like this:

```
h, cp, p, U, mu = hg.pettitt_test(x, 0.05)
```

User can be plot their results by follow:

```
mn = 0
mx = len(data)

loc = result.cp
mu1 = result.avg.mu1
mu2 = result.avg.mu2

plt.figure(figsize=(16,6))
plt.plot(data, label="Observation")
plt.hlines(mu1, xmin=mn, xmax=loc, linestyles='--', colors='orange',lw=1.5, label='mu1 : ' +
plt.hlines(mu2, xmin=loc, xmax=mx, linestyles='--', colors='g', lw=1.5, label='mu2 : ' + str
plt.axvline(x=loc, linestyle='-.' , color='red', lw=1.5, label='Change point : '+ str(loc) +

plt.title('Title')
plt.xlabel('X')
plt.ylabel('Y')
plt.legend(loc='upper right')

plt.savefig("F:/aaaaaa.jpg", dpi=600)
```

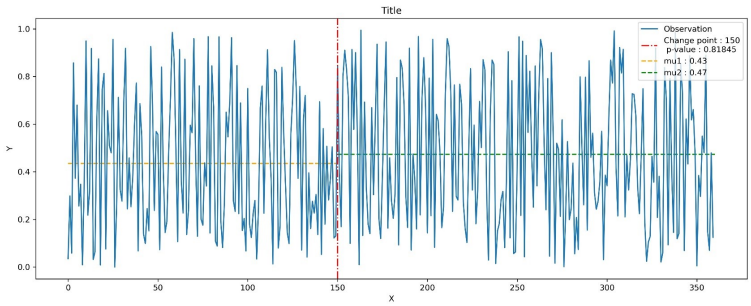

Figure 1: Homogeneity result plot

Acknowledgements

This work is done under the project *"Python packages/tools development for environmental research".*

References