Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *doi: 10.18637/jss.v000.i00*

A **Python/Zig** optimized and customizable implementation for the ρ_{DCCA} and DMC_x^2 methods

Fernando Ferraz Ribeiro
Universidade Federal da Bahia

Gilney Figueira Zebende
Universidade Estadual de Feira de Santana

Abstract

This paper presents tha **Zebende**, a Python package written in Python and Zig, that calculates the DFA, $DCCA$ ρ_{DCCA} and the DMC_x^2. The package presents an optimized algorithm that significantly improves the calculations speed. A comparison with other packages that calculates the .The package is also the first to implement the DMC_x^2 coefficient for any number of series.

Keywords: ρ_{DCCA}, DMC_x^2, optimization, Python, Zig.

1. introduction

The ρ_{DCCA} (Zebende 2011) is a widely used coefficient that measures the cross-correlation between tow non-stationary time series. It´s an extension of the Detrended Fluctuation Analysis (DFA) (Peng, Buldyrev, Havlin, Simons, Stanley, and Goldberger 1994) and the Detrended Cross-correlation Analysis ($DCCA$) (Podobnik and Stanley 2008): while the DFA calculates the self-affinity and long-memory properties of a time series data, and the $DCCA$ analyses power-law cross correlations between two different non-stationarity time series, the ρ_{DCCA} coefficient quantifies this cross-correlation in simple values ranging from -1 to 1, where -1 indicates a perfect anti-correlation between the series, 1 a perfect correlation and zero (0) no correlation at all.

The Detrended Multiple Cross-Correlation Coefficient (Zebende and Silva 2018) (DMC_x^2) is a generalization of the ρ_{DCCA} coefficient that correlates one time series (dependent variable) a number of time series (independent variables). The DMC_x^2 values ranges from zero (0), indicating no correlation to 1, meaning perfect correlation or anti-correlation between the dependent and the independent variables.

This paper presents the **Zebende** Python package, an implementation of the DFA, $DCCA$, ρ_{DCCCA}, DMC_x^2 and utility functions related to the methods. In section 2 the steps for calculating the ρ_{DCCCA} and DMC_x^2 are presented and discussed. Section 3 shows how this library was implemented, the optimization technics and the recommended steps to use the library. In Section 4 the **Zebende** package is compared with other packages for Python and R that calculates the ρ_{DCCA} in terms of performance and usability, leading to the conclusions in 5.

2. Algorithms of the coefficients

The algorithms that calculates the ρ_{DCCA} uses the DFA and the $DCCA$ steps. The DMC_x^2 coefficient uses the ρ_{DCCA} coefficient and, consequently, also embraces the DFA and the $DCCA$. The DFA method is described in six steps:

1. Taking a time series $\{x_i\}$ with i ranging from 1 to N, the integrated series X_k is calculated by $X_k = \sum_{i=1}^{k} [x_i - \langle x \rangle]$ with k also ranging from 1 to N;

2. the X_k series is divided in $N - n$ boxes of size n(time scale), each box containing $n + 1$ values, starting in i up to $i + n$;

3. for each box, a polynomial (usually of degree 1) best fit is calculated, getting $\widetilde{X}_{k,i}$ with $i \leq k \leq (i + n)$ (detrended values);

4. in each box is calculated: $f_{DFA}^2(n, i) = \frac{1}{1+n} \sum_{k=i}^{i+n} (X_k - \widetilde{X}_{k,i})^2$

5. for all the boxes of a time scale, the DFA is calculated as:

   $F_{DFA}(n) = \sqrt{\frac{1}{N-n} \sum_{i=1}^{N-n} f_{DFA}^2(n, i)}$;

6. for a number of different timescales (n), with possible values $4 \leq n \leq \frac{N}{4}$ the F_{DFA} function is calculated to find a relation among $F_{DFA} \times n$

The $DCCA$ method is very similar to the DFA calculations. With the difference of analyzing two series while the DFA evaluate properties of a single time series. It´s also a six steps process:

1. Taking two time series with the same length $\{x\alpha_i\}$ and $\{x\beta_i\}$ with i ranging from 1 to N, the integrated series $X\alpha_k$ and $X\beta_k$ are calculated by $X_k = \sum_{i=1}^{k} [x_i - \langle x \rangle]$ for each series, with k also ranging from 1 to N;

2. $X\alpha_k$ and $X\beta_k$ series are divided in $N - n$ boxes of size n(time scale), each box containing $n + 1$ values, starting in i up to $i + n$;

3. for each box, a polynomial (usually of degree 1) best fit is calculated, getting $\widetilde{X\alpha}_{k,i}$ and $\widetilde{X\beta}_{k,i}$, for series $\{x\alpha_i\}$ and $\{x\beta_i\}$ respectively, with $i \leq k \leq (i + n)$ (detrended values);

4. in each box is calculated: $f_{DCCA}^2(n, i) = \frac{1}{1+n} \sum_{k=i}^{i+n} (X\alpha_k - \widetilde{X\alpha}_{k,i}) \times (X\beta_k - \widetilde{X\beta}_{k,i})$

5. for all the boxes of a time scale, the $DCCA$ is calculated as:

   $F^2_{DCCA}(n) = \frac{1}{N-n}\sum_{i=1}^{N-n} f^2_{DCCA}(n,i)$;

6. for a number of different timescales (n), with possible values $4 \leq n \leq \frac{N}{4}$ the F^2_{DCCA} function is calculated to find a relation among $F^2_{DCCA} \times n$

Comparing the algorithms, the first three are basically identical, the only difference is that the $DCCA$ method apply those steps to two series. The step four of the DFA is essentially an application of the variance calculation and the equivalent step of the $DCCA$ is a covariance between the two series. Step five calculates the square root of the mean of the variances calculated in in each box for the DFA, in the $DCCA$, the mean of the covariances calculated for each box is calculated in stead. The last step,in both cases, is more a reminder to repeat the respective previous operations for a number of difference time scales.

The ρ_{DCCA} is measured using Eq. 1. Considering the relation between DFA and variance and $DCCA$ and covariance, the ρ_{DCCA} resembles Pearson correlation for a time scale n.

$$\rho_{DCCA}(n) = \frac{F^2_{DCCA\ (x\alpha,\ x\beta)}(n)}{F_{DFA\ (x\alpha)}(n) \times F_{DFA\ (x\beta)}(n)} \tag{1}$$

The DMC^2_x is a generalization of the ρ_{DCCA} that calculates the correlation between one time-series $\{Y\}$, as the dependent variable, and a number j of time-series $\{X_1\}$, $\{X_2\}$, $\{X_3\}$, $\ldots$, $\{X_j\}$ defined as independent variables. The coefficient is expressed mathematically as:

$$DMC^2_x \equiv \rho_{Y,X_i}(n)^T \times \rho^{-1}(n) \times \rho_{Y,X_i}(n) \tag{2}$$

In Eq. 2, $\rho^{-1}(n)$ represent the inverse of a matrix populated by all possible combinations of ρ_{DCCA} between independent variables. In Eq. 3, value $\rho_{X_1,X_2}(n)$, for instance, is the ρ_{DCCA} for independent variables X_1 and X_2 calculated with time scale n, occupying position ρ_{12} of the matrix. Two fundamental characteristics: the first is that the main diagonal values are all ones, since it's position in the matrix denotes the calculation of a correlation between a series and itself. Second, the matrix is symmetric in relation to the main diagonal, as the ρ_{DCCA} is evaluated a commutative expression.

$$\rho^{-1}(n) = \begin{pmatrix} 1 & \rho_{X_1,X_2}(n) & \rho_{X_1,X_3}(n) & \ldots & \rho_{X_1,X_j}(n) \\ \rho_{X_2,X_1}(n) & 1 & \rho_{X_2,X_3}(n) & \ldots & \rho_{X_2,X_j}(n) \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \rho_{X_j,X_1}(n) & \rho_{X_j,X_2}(n) & \rho_{X_j,X_3}(n) & \ldots & 1 \end{pmatrix}^{-1} \tag{3}$$

At last Eq. 4 represent the transposed vector of the $\rho_{Y,X_i}(n)$ between the depended variable $\{Y\}$ and each $\{X_i\}$ independent variable for a given time scale n.

$$\rho_{Y,X_i}(n)^T = [\rho_{Y,X_1}(n), \rho_{Y,X_2}(n), \cdots, \rho_{Y,X_j}(n)] \tag{4}$$

The ρ_{DCCA} and the DMC^2_x should be evaluated in a number of time scales (n) to analyze the characteristics of each coefficient.

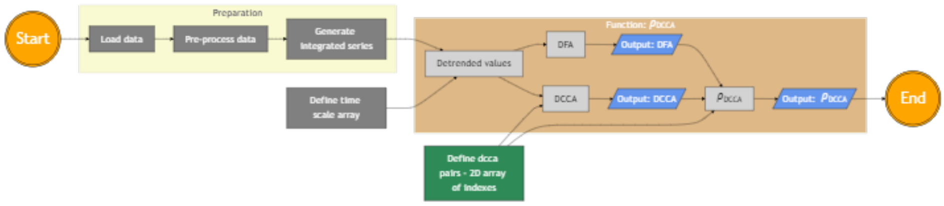

Figure 1: Flowchart for the ρ_{DCCA} package usage

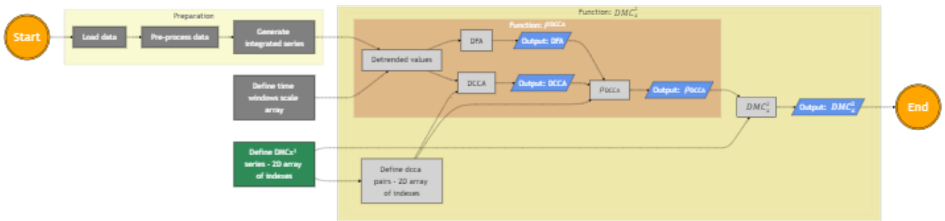

Figure 2: Flowchart for the DMC_x^2 package usage

3. Zebende package: implementation and optimization

4. Results

5. Summary and discussion

References

Peng CK, Buldyrev SV, Havlin S, Simons M, Stanley HE, Goldberger AL (1994). "Mosaic Organization of DNA Nucleotides." **49**(2), 1685–1689.

Podobnik B, Stanley HE (2008). "Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series." *Physical Review Letters*, **100**(8). ISSN 00319007. doi:10.1103/PhysRevLett.100.084102. 0709.0281.

Zebende GF (2011). "DCCA cross-correlation coefficient: Quantifying level of cross-correlation." *Physica A: Statistical Mechanics and its Applications*, **390**(4), 614–618. ISSN 03784371. doi:10.1016/j.physa.2010.10.022. URL http://dx.doi.org/10.1016/j.physa.2010.10.022.

Zebende GF, Silva AM (2018). "Detrended Multiple Cross-Correlation Coefficient." *Physica A*, **510**, 91–97. ISSN 0378-4371. doi:10.1016/j.physa.2018.06.119.

Affiliation:

Fernando Ferraz Ribeiro
Universidade Federal da Bahia
Faculty of Achitecture
Universität Innsbruck
Universitätsstr. 15
6020 Innsbruck, Austria
E-mail: fernando.ribeiro@ufba.br
and
Centro Universitário Senai-Cimatec

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd