Anomaly Detection Library

Kapil Thakkar and Reshma Kumari

27th May, 2016

Contents

Chapter 1

Window Based Correlation

1.1 Introduction

This technique is basically applied on two time-series. Let's say we have two time series as series1 and series2. So, in this method, we first find correlation at various lags between these two time series. User can specify minimum and maximum lag to consider. So, for all those values, we find correlation values.

After finding correlation values at all lags, we consider that lag at which correlation value is higher among all previously calulated correlation values at all lags. Let's say that lag be "x". So, depending upon that "x", we shift series1 or series2. If "x" is positive, we move series2 by "x" units and if it is negative than we shift series1 by $|x|$ units.

Now, we are ready to apply window correlation. Take window value,"w" as input. First window will be from 1st element to w'th elementof both the time series after aligning by lag "x". Find correlation for this window between two time-series and save it in an array. Now, slide window by "w" elements and calculate correlation value again and so on. Now, we have correlation values at multiple windows.

Now, let's say both the series should have been positively correlated. So, what we do is, we choose threshold by MAD test if not provided to us, and find all correlation values which are below that threshold and report all those windows as anomaly.

1.2 Related Functions

1.2.1 correlation(arr1, arr2, maxlag, pos, neg)

This function calculates correlation between arr1 and arr2 at all possible lags between -maxlag to +maxlag, as specified by pos and neg parameters.

- Input Parameters
    1. arr1 *(list)* : Input series 1 as a list of float values
    2. arr2 *(list)* : Input series 2 as a list of float values
    3. maxlag *(int)* : maximum (maxlag) and minimum (-maxlag) lag to consider while calculating correlation betweem arr1 and arr2

4. pos *(int, 1 or 0)* : To consider positive lag or not, i.e. 1 to maxlag
5. neg *(int, 1 or 0)* : To consider negative lag or not, i.e. -maxlag to -1

- Output *(list)* :
  Returns list of tuples of the form

(lag, correlation value at this lag)

1.2.2 getMaxCorr(arar1,positive_correlation)

This function takes list of tuples of the form (lag, correlation value at this lag) as input. Returns lag value at which correlation value is maximum if positive_correlation is True, and returns lag at which correlation value is minimum if positive_correlation is False.

Basically, if both the series are positively correlated than we will be interested in maximum positive correlation or if both series are negatively correlated than we will be interested in minimum negative correlation, which is specified by positive_correlation parameter.

- Input Parameters

1. arr1 *(list)* : list of tuples of the form
   (lag, correlation value at this lag)
   i.e. correlation values at various lags
2. positive_correlation *(boolean, "True" or "False")* :
   - True: If value of this parameter is True than it will return lag at which correlation value if maximum (positive)
   - False: If value of this parameter is False than it will return lag at which correlation value if minimum (negative)

- Output *(Tuple)* :
  returns single tuple of the form (lag,correlation value at this lag), i.e. lag at which optimum correlation value is found along with correlation value.

1.2.3 correlationAtLag(series1, series2, lag, window_size)

This function fisrt aligns two series by given lag. If lag is positive than it shifts start of series2 else start of series1. After aligning both the series according to lag, this function calculates correlation between both series at all windows.

window_size states size of the window. So, we will start with first window taking first window_size elements from each series and will calculate correlation. We will save this correlation value in list and will slide to next window. Next window will start after window_size elements. In such a way, we calculate, correlation at all windows and return the list of correlation values.

- Input Parameters

1. series1 *(list)* : Input series 1 as a list of float values

2. series2 *(list)* : Input series 2 as a list of float values
3. lag *(int)* : lag at which series needs to be adjusted as explained above
4. window_size *(int)* : window size to be considered

- Output *(list)* :
  Returns list of correlation values (of float type) for all windows calculated at given lag

1.2.4 WindowCorrelationWithConstantLag(arr1, arr2, window_size,maxlag, positive_correlation, pos, neg)

This is sort of driver function, which will call above 3 functions. This function will first get lag at which series needs to be adjusted. Than using this lag, it will calculate correlation values at all windows and will return it.

- Input Parameters

  1. arr1 *(list)* : Input series 1 as a list of float values
  2. arr2 *(list)* : Input series 2 as a list of float values
  3. window_size *(int)* : window size to be considered while calculating window correlation
  4. maxlag *(int)* : maximum (maxlag) and minimum (-maxlag) lag to consider while calculating correlation betweem arr1 and arr2, to align both the series
  5. positive_correlation *(boolean, "True" or "False")* :
     - True: This suggest that both the series are positively correlated
     - False: This suggest that both the series are negatively correlated
  6. pos *(int, 1 or 0)* : If value of this parameter is True than we will consider positive values for lag, i.e. 1 to +maxlag to align both the series initially
  7. neg *(int, 1 or 0)* : If value of this parameter is True than we will consider negative values for lag, i.e. -maxlag to -1 to align both the series initially

- Output *(list)* :
  List of tuples of the form (lag,array) Where lag is lag value for which whole series is shifted and then at that lag, we have calculated correlation for all window. Correlation value for all windows is stored in array.

1.2.5 anomaliesFromWindowCorrelationWithConstantlag(arr1, arr2, window_size=15,maxlag=15, positive_correlation=True, pos=1, neg=1, default_threshold = True, threshold = 0):

This is main function of this method. This is driver of whole method. Using previously stated methods, it will first gather correlation values at different windows. Than depending upon which type of threshold is to be used, it will filter

out anomalies. If default threshold is to be used, than it will be caulated using MAD test on the correlation values at each window, else threshold provided by user will be used.

Correlation values not satisying threshold will be reported along with the date range of that window.

- Input Parameters

  1. arr1 *(list)* : Input series 1 as a list of tuples of the from (date,value)
  2. arr2 *(list)* : Input series 2 as a list of tuples of the from (date,value)
  3. window_size *(int)* : window size to be considered while calculating window correlation
  4. maxlag *(int)* : maximum (maxlag) and minimum (-maxlag) lag to consider while calculating correlation betweem arr1 and arr2, to align both the series
  5. positive_correlation *(boolean, "True" or "False")* :
     - True: This suggest that both the series are positively correlated
     - False: This suggest that both the series are negatively correlated
  6. pos *(int, 1 or 0)* : If value of this parameter is True than we will consider positive values for lag, i.e. 1 to +maxlag to align both the series initially
  7. neg *(int, 1 or 0)* : If value of this parameter is True than we will consider negative values for lag, i.e. -maxlag to -1 to align both the series initially
  8. default_threshold *(boolean, "True" or "False")* : whether to use default threshold or not. If True, default threshold will be used using MAD test on calculated correlation values for all windows.
  9. threshold *(float)* : if default_threshold is False, than this user provided threshold will be used.

- Output *(list)* :
  This function filter out anomalies and returns them. This function returns List of tuples of the form
  (start_date,end_date,correlation_value),
  where (start_date, end_date) specifies range of the window and correlation_value if value of correlation of that window

1.3 Description

Putting all together, here is the summary:

Function "WindowCorrelationWithConstantLag", first makes use of "correlation" function, to calculate correlation values at all lags to find out at which lag it needs to be align. Result of "correlation" function is passed to "getMaxCorr" function. Which will return lag at which optimum value of correlation is present. This output will be used by "correlationAtLag" function, to caluclate correlation at all windows after aligning both series by input lag. So, in this way

"WindowCorrelationWithConstantLag" combines these three functions and returns correlation value at each window.

Function "anomaliesFromWindowCorrelationWithConstantlag" is the main driver. This function calls "WindowCorrelationWithConstantLag" and gets the correlation values at all windows and filters out anomalies (either using threshold calculated by MAD test or by user provided threshold) and returns them in the format of (start_date,end_date,correlation_value), where (start_date, end_date) specifies range of the window and correlation_value if value of correlation of that window.