Israel Diego (UMID: israeldi)
Stats 504
December 16th

Fractional Differencing of Internet Traffic Data

The CAIDA darkspace internet traffic data is a time-series dataset measuring the number of UDP and TCP packets on a 1-minute time-scale for April 2012. We conduct our analysis on UDP and TCP time-series and attempt to achieve stationarity via fractional differencing.

From the darkspace dataset, we calculate: total packets, number of unique sources, logged number of UDP packets, and logged number of TCP packets for April 1st, 2012. The difference between UDP and TCP, UDP is used for applications like video conferencing, while TCP is used for handling most email and web pages. Each of these variables comprises a time series with 1440 values, corresponding to the number of minutes in a day.

If a time-series appears non-stationary, a common approach is to apply first-differences in order to make it stationary. If the original series is $y_1, y_2, \ldots, y_n$, then the first-differenced series is $y_2 - y_1, y_3 - y_2, \ldots, y_n - y_{n-1}$. Similarly, the second-differenced series would be the difference of the first-differences,

$$(y_3 - y_2) - (y_2 - y_1), \ldots, (y_n - y_{n-1}) - (y_{n-1} - y_{n-2}) = (y_3 - 2y_2 + y_1), \ldots, (y_n - 2y_{n-1} + y_{n-2})$$

We can formulate this process more succinctly by using backshift/lag operator B. Applying the operator to a vector y_t, produces $By_t = y_{t-1}$. Then, the first-differenced series may be written as,

$$(1 - B)y_t = y_t - By_t = y_t - y_{t-1}$$

Likewise, the second-differenced series may be written as,

$$(1 - B)^2 y_t = y_t - 2By_t + B^2 y_t = y_t - 2y_{t-1} + y_{t-2}$$

Thus, we can generalize the integer n^{th}-order differences by the binomial formula, $(1 - B)^n = \sum_{k=0}^{n} \binom{n}{k}(-B)^k$. Fractional-differencing allows us to extend this approach by letting the exponent be any real number. Using the taylor series expansion of the binomial formula about $B = 0$, and letting $n = d$, where $d \in \Re$, yields,

$$(1 - B)^d = 1 - dB + \frac{d(d-1)}{2!}B^2 - \frac{d(d-1)(d-2)}{3!}B^3 + \cdots = \sum_{k=0}^{\infty} \binom{d}{k}(-B)^k$$

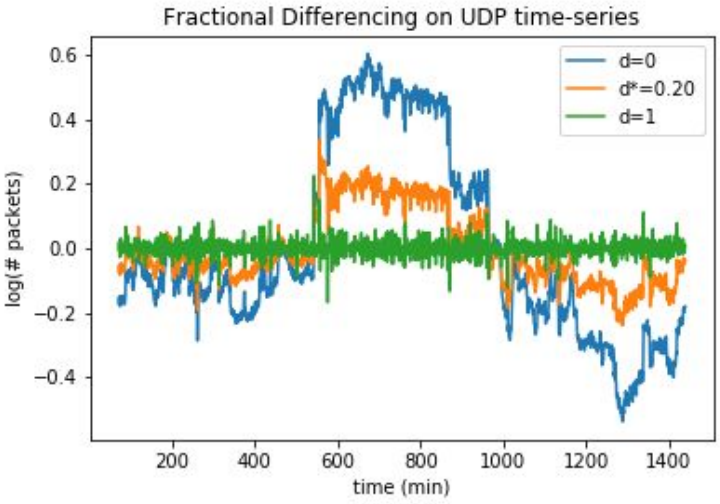
Now we have a means for applying fractional-differencing to time-series data. We approximate the taylor series by setting a threshold for the smallest term. For our analysis, we set a threshold of 0.001, such that we retain all coefficients of the lag operator that are bigger than this threshold. We compute fractional differences based on a fixed-width window, such that we always use the same number of weights at each time step according to our approximation.

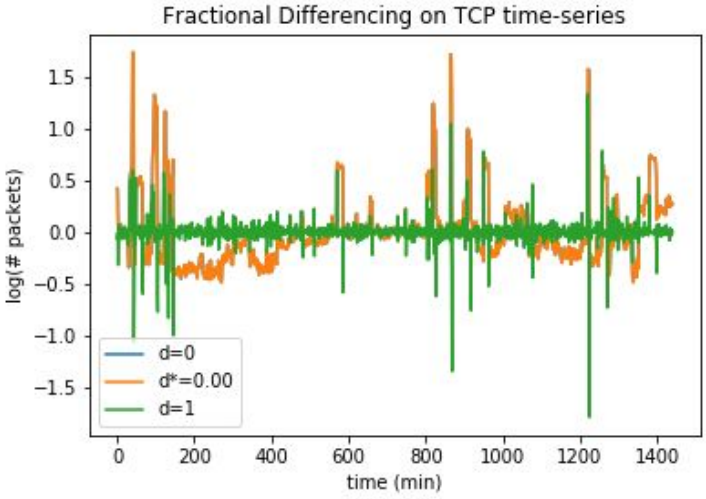
Although we have defined the framework for fractional differencing, we seek to find the optimal level of differencing necessary to achieve a stationary time-series and avoid over-differencing. A popular approach to measure stationarity (or non-stationarity) is by Augmented Dickey Fuller (ADF) test. The ADF test, tests the null-hypothesis that our time-series has a unit-root. If we fail to reject, then the time-series is non-stationary.

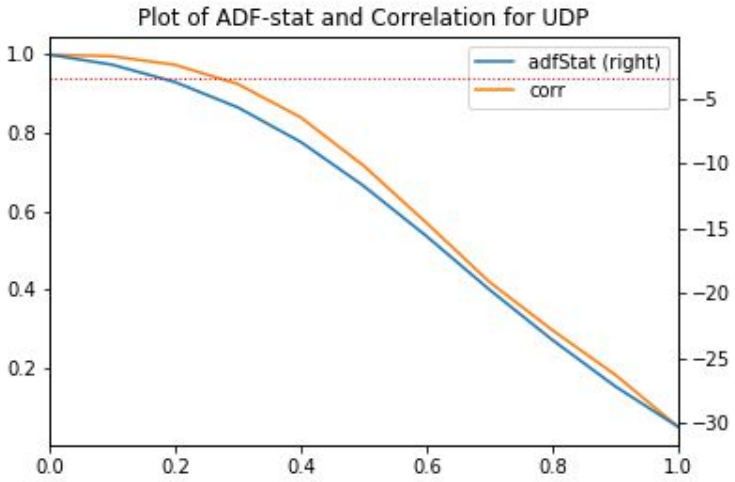
In **Figure 1**, the bottom-left plot shows the ADF statistic plotted against the order of differencing d. We also plot the correlation of the d-differenced time-series with respect to the original series, which allows us to measure the amount of information retained after differencing. For UDP, we note that first-order differencing produces an ADF statistic of -30.31, which is large enough to conclude there is no unit root, and low correlation ≈ 0.05. Although we have a stationary time-series after first-differencing, we have lost almost all memory/predictability from our data. The horizontal dashed line is the critical value of the ADF test (-3.74) at the 1% level that is necessary to achieve stationarity. By intersecting the dashed line and ADF statistic (blue line), we find our optimal choice of d is about 0.2 and the corresponding correlation is about 0.97. This implies that we have achieved stationarity while retaining most of the information from the original series. In order to approximate the taylor series, we required 72 terms at each time step. The top-left plot shows the original series, the optimal d-differenced series, and the first-differenced series. We see that indeed, the fractionally-differenced series looks stationary. When conducting the same analysis on TCP (right plots), we conclude that the original series with no differencing is already stationary since the corresponding ADF statistic is significant at the 1% level. If we were to apply first-differencing to the TCP series, the correlation would drop to 0.2 and we would lose a good amount of information.

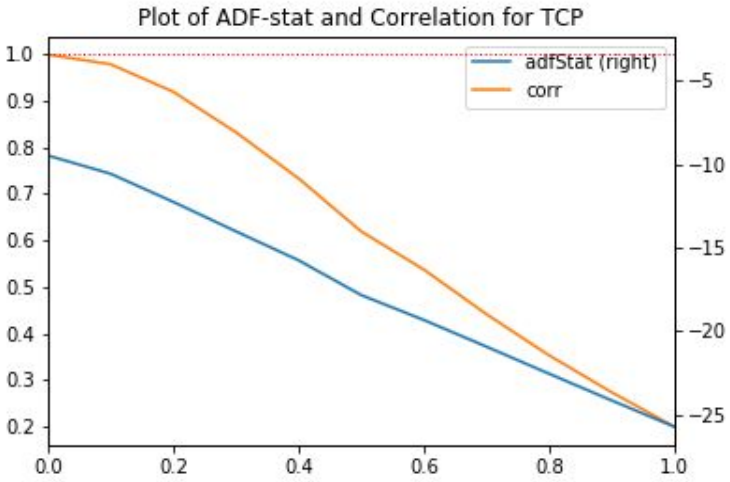
We find that fractional differencing is capable of producing a stationary time series while retaining more information than first-order differencing. This can be advantageous in numerous statistical applications such as modeling and forecasting.

Figure 1









Order-difference d

Order-difference d