Statistics Trading:Distance approach Part 1

2019.03.30

Gatev, Goetzmann and Rouwenhorst

1. Their study performed on all liquid U.S. stocks from the CRSP daily files from 1962 to 2002. First
2. A cumulative total return P_{it} is constructed for each stock i and normalized to the first day of a 12 months formation period.
3. With n stocks under consideration, the sum of Euclidean square distance(SSD) for the price time series of $n(n-1)/2$ is calculated

Gatev, Goetzmann and Rouwenhorst

Implementation:

1. Select the top 20 pairs with minimum historic distance metric are considered in a subsequent six months trading period. Prices are normalized again to the first day of the trading period.
2. Trades are opened when the spread diverges by more than two historical standard deviation σ ad closed upon mean reversion, at the end of trading period, or upon delisting.

Gatev, Goetzmann and Rouwenhorst

The advantage of GGR method

1. The methodology is relatively clear.
2. Easy to implement
3. Robust to data snooping

The disadvantage of GGR method

1. The choice of Euclidean squared distance for identifying pairs is analytically suboptimal.

Gatev, Goetzmann and Rouwenhorst

The idea measure should be able to identify the pair with

1. The spread should exhibit high variance
2. The spread should be strongly mean-reverting.

These two attributes generate a high number of round trip trades with high profits per trade.

Spread Variance

Empirical spread variance $V(P_{it} - P_{jt})$ can be expressed as follows:

$$V(P_{it} - P_{jt}) = \frac{1}{T}\sum_{t=1}^{T}(P_{it} - P_{jt})^2 - (\frac{1}{T}\sum_{t=1}^{T}(P_{it} - P_{jt}))^2$$

We can solve the average sum of squared distances for the formation period:

$$\bar{SSD}_{ijt} = \frac{1}{T}\sum_{t=1}^{T}(P_{it} - P_{jt})^2 = V(P_{it} - P_{jt}) + (\frac{1}{T}\sum_{t=1}^{T}(P_{it} - P_{jt}))^2$$

Spread Variance

First of all, it is trivial to see that an ideal pair in the sense of GGR with zero squared distance has a spread of zero and thus produces no profits.
The decomposition of SSD shows that there are two effects influence ranking of GGR.

Mean reversion

GGR interpret the pairs price time series as cointegrated in the sense of Bossaerts(1988).
However, GGR perform no cointegration testing on their indentified pairs. As such, the high correlation may well be spurious, since high correlation is not related to a cointegration relationship.

better selection measure

1. pairs exhibiting the lowers drift in spread mean should be identified.
2. of these pair, the ones with the highest spread variance are retained and tested for cointegration while controlling the familywise error rate as in Cummins and Bucca(2012).