What Can Behavioral Biometrics Tell Us About the Stock Market? (Part I)

By
John C. Checco, CISSP

(rev. 0.3)

Introduction

The goal for an economist is similar to any scientist: to define a predictive analysis of what we see and experience everyday. Instead of looking for new external factors that affect the stock market, as derivative do; we have utilized our expertise in behavioral biometric technology to find time-delayed patterns of consistency correlating company stock activity across markets.

What Are Behavioral Biometrics?

The idea behind biometrics is to identify one person from another given a physical or behavioral characteristic that offers a high degree of non-repudiation. The classification of physical and behavioral is important in this discussion because the underlying requirements of what makes a good biometric could not be more different. Physical biometrics is what one commonly sees in the movies – fingerprint, handprint, iris recognition, retinal scans, et al. – it measures precise data points against a known template.

Behavioral biometrics uses pattern recognition against captured samples of user actions. Some well-known types of behavioral biometrics are voice recognition, handwriting recognition and keystroke dynamics. Whereas physical biometrics produce a pass/fail result; behavioral biometrics provide a confidence measurement which then must be accepted or rejected based on other dependent thresholds.

Why Behavioral Biometrics?

The concept behind behavioral biometrics ascertains that variations in signatures are not only allowed, but expected, because the behavior being measured is **known** to be pattern specific. bioChec™ uses a patent-pending keystroke biometric algorithm that does the two things expected of a behavioral biometric engine: create a pattern template from a known set of N samples; and determine the confidence level of a signature against an existing template.

The item of interest here is the template creation. For traditional uses, behavioral biometric implementations create a very precise matching pattern for those data points which are known to exhibit consistent patterns. The bioChec™ template creation algorithm diverges from this tradition in two ways:

1. **CONSISTENCY:** Never assume that samples generating a template are consistent. For each template created, a strength factor is calculated that determines how consistent the sample set is – i.e. how consistent is the user's typing. From a traditional standpoint, disregarding this assumption creates a higher "Failure to Enroll Rate" – but also allows for higher confidence thresholds for those sample sets which create strong templates.

2. **DYNAMICS:** Never assume that behavior can be defined by a static sample set. Although it is very rare that someone's fingerprint, retina or other physical biometric may change, most biometric algorithms consider this the anomaly. While we can agree that this is the right approach for physical biometrics, we believe that a behavioral biometric is *meant* to change over time. This means we must always consider samples over time to

generate and regenerate usable patterns.

So, from our perspective, a good behavioral biometric template consists of relatively consistent, but slightly changing samples over time. For our implementation, we constantly recalculate the template pattern (and strength) with new data over time. In our experience with user keystroke behavior, we have seen firsthand the significant increase in template strengths when the sample set goes from 10-20 samples (used by other behavioral biometric implementations) to 1000-2000 samples in our implementation.

Where Biometrics Meets the Stock Market

The application of consistency and dynamics in template creation can be further extrapolated to not only create good patterns; but even to verify that a pattern exists (albeit within a very narrow set of parameters).

Why Keystroke Biometrics?

Keystroke biometrics is unique from other behavioral biometrics in that it measures time within a small set of known actions (key press, key release) against a known set of variants (keyboard keys) which can then be aggregated into computational pieces of flight time (the time it takes to get from one key to the next) and dwell time (the time it takes from a key press to its matching release).

In other words, keystroke biometrics contains the necessary elements needed to measure the stock market: a limited set of actions (rise/fall of stock quote) against a known set of variants (stock symbols) across a controlled timeframe.

What Exactly Are We Measuring?

What we were trying to accomplish with this phase-1 stock quote experiment is determine if the bioChec™ keystroke template creation algorithm can be used to deduce a time-delayed consistency among heavily repeated, but seemingly arbitrary, data points; hence the stock quotes. Quotes are generated consistently every business day, along with specific indices. We are using the bioChec™ algorithm to process this data as if it were a known behavioral biometric -- and from looking at the template "strength" whether there are real time-delayed consistency patterns.

For example, if the Dow Jones Industrial Average rises 1% on a Tuesday (close of business), we may find that ACME stock consistently rises between 0.15 - 0.25% by Wednesday (close of business), given a certain margin of variation. This is, by far, a most simple example to illustrate what relationships could exist.

The Phase-I Experiment

The simple question we are asking is:

> *Removing all market boundaries, are there stocks in one market that innately follow stocks or indices in other markets in a time-delayed reaction?*

The idea is a bit naïve: does one company's value depend on another? The obvious answer is "yes" if the companies are in some type of professional relationship. But this fact actually helps us validate our findings. There are some correlations that are to be expected. Thus if those correlations are found to be strong; what can be surmised from other strong yet non-trivial correlations found?

Factors and Controls

How does one attack a problem with such a broad hypothesis? The raw computing power needed to run calculations far exceeded our capabilities. Concessions were made with respect to data inputs and candidates for further processing. For a sample set of N stocks with a time delay range up to D days needs an extraordinary number of raw calculations nightly to generate the following number of pattern data points: `"(combination of N elements evaluated 2 at a time)*2*D".` As a result, the following factors and controls were created:

<u>Target Data Set:</u>
Given that the bioChec™ algorithm generates huge amounts of complex calculations per data element, the implementation of such a broad concept would be computationally impossible. So we needed to limit the experiment in a controlled fashion; to use data that is logically sound and could generate repeatable results. We decided to reduce the quotes used to 20 indices and the approximately 675 stocks that make up the DJIA, ETF, NASDAQ 100, and S&P 500.

<u>Time Differential Control:</u>
An initial set of time delays of 1–5 days was chosen to measure. We could have included same-day correlations; but we assumed there would have been too many "high" expected correlations to disseminate any new correlations. We also chose an upper time delay of 5 days to limit the noise generated in calculations; the probability of any significant correlations beyond day 5 would be highly unlikely given the limited dataset chosen.

<u>Correlation Limitations:</u>
Evaluations of the "Top 100" correlations were run periodically to track several items: (1) find the controlled relationships, (2) determine the variations in the correlations themselves, and (3) how correlations varied over time. Tracking these items can lend to further derivations which removed known external factors; such as mergers and acquisitions and other news items.

<u>Directionality Limitations:</u>
After some initial runs, we found that correlations between two stocks could spike in opposite directions occasionally; thus minimizing the effect of the pattern strength built up to that spike. A decision was made to include an additional factor of "directionality" – whether the daily calculated pattern element would fall into a direct relationship bucket or an inverse relationship bucket. The net result is that for every stock pair; there are 2 possible buckets where one bucket was more significant than the other in defining the overall correlation. This also helped to rule out non-trends where both directions had significantly near strengths.

Control Validation:
To confirm the new implementation of this algorithm would indeed be able to identify trends, a series of hand-crafted highly correlated pairs of dummy stocks were input and the calculations were verified to be accurate. Conversely, known bad data was explicitly introduced to verify that "non-trends" were not falsely correlated.

Processing Details

To process this initial dataset of indices and stocks: we used a dual-processor i686 Red Hat Linux System with Oracle 8i and 1GB of RAM. It took about 3-4 hours to run about 1.5M - 2.5M calculations nightly, and another 2 hours to apply consistency rule sets (a.k.a. correlation runs). These correlation runs picked the top 100 correlations calculated for that day's closing data based on strongest template strength.

We backfilled the database with daily data and calculations from January 1, 2002 -- which took about 4 months to catch up to normal nightly calculations. The last data feed was sometime in March 17, 2005.

Post March 17, 2005, some processing was done against all the "Top 100" picks to match expected time-delayed dependent rates change with actual rate change for that pick's target date. From these values, we were able to see how close our consistency ratings were against reality.

Raw Results

Within the 39 months of data gathered, there were:

- 532,647 quotes read,
- 824 days of quotes
- 1,586,576,690 total calculations
- 1,925,457 average calculations daily

For the measurement from the close of one stock to the close of another stock, there were:

- 7,140,306 pattern data points generated
- sample size for each data point ranged from 1 to 626 samples (297 on average)

From these measurements, correlation calculations were run

- Each correlation run generated a "Top 100" list for that run date
  - The 100 strongest **direct** relationships
  - The 100 strongest **inverse** relationships
- 147 correlation runs between 6/11/2004 and 3/17/2005
- 28,714 correlations total.

As expected, raw measurements ranged wildly (because not

Technical Sidebar:

In the course of this experiment, we uncovered a major Oracle bug where the MERGE statement cannot resolve calls to an OBJECT's member function directly. Rather, the function must be wrapped by a global static function call which increases overhead and execution costs. The bug was reported to Oracle in April of 2004, but has yet to be fixed as of version 10g. (Refer to TAR 3872782.999 – "ERROR COMPILING A STORED PROC WITH USING "MERGE" WITH A MEMBER FUNCTION" on Oracle's MetaLink site.)

everything can be deemed a pattern)

- standard deviations of percent change ranged from 0 to 7151 (26 on average)
- error percentages ranged from -6,099,546,194 to 36,957,762,267 (7972 on average)

Correlation Results

From the "Top 100" reports generated, we could now limit higher level analysis to data points which exhibited the highest consistency ratings for a given period of time. From this reduced data set we found the following highlights:

- highest correlation (from "Top 100" correlation runs)

Symbol 1	LU **LUCENT TECH INC** Market_Cap=**11.609 B**
Close Date	2004-11-10
Breadth	1 Day
Symbol N	SHY **ISHARES LEHMAN 1** Market_Cap=(<100M)
Relationship	Inverse
Expected Close	81.76
Actual Close	81.69
Sample Set	**213**
Std Dev	**0.19**
Percent Error	855.47

- lowest correlation (from "Top 100" correlation runs)

Symbol 1	APC **ANADARKO PETROLEU** Market_Cap= **18.643 B**
Close Date	2004-06-19
Breadth	3 Days
Symbol N	JDSU **JDS UNIPHASE CP** Market_Cap= **2.283 B**
Relationship	Inverse
Expected Close	3.11
Actual Close	3.58
Sample Set	37
Std Dev	44.34
Percent Error	-933.78

- most "Top 100" correlation picks (from 147 correlation runs)

Symbol 1	BMY **BRISTOL MEYERS SQI** Market_Cap= **49.673 B**
Correlation Runs	**80**

Breadth	1 Day
Symbol N	SHY **ISHARES LEHMAN 1** Market_Cap=(<100M)
Relationship	Direct
Std Dev	0.056

- most consistent correlation (with minimum correlation runs > 10)

Symbol 1	CR **CRANE CO** Market_Cap= **1.648 B**
Correlation Runs	16
Breadth	2 Days
Symbol N	SHY **ISHARES LEHMAN 1** Market_Cap=(<100M)
Relationship	Direct
Std Dev	**0.034**

Phase-I Conclusions

Initially we were trying to find any time-delay pattern, but our conclusions suggest that the highest percentages of correlation were found at companies with the lowest levels of capitalization.

Relatively speaking, since all the companies selected were part of one index or another, even the smallest companies were fairly large and the highest percentage of correlation was fairly small.

It is important to note the template strengths calculated for even the strongest correlations were well below an acceptable level if this were biometric data.

<u>Top 100 Reports:</u>

The "Top 100" correlation runs were specifically designed to minimize the amount of calculations needed to be analyzed post mortem. Given that for any day there were on average 2 million possible calculations to consider, 100 per day for 824 days deemed to be a reasonable sampling. From the post mortem analysis, 2 major findings emerged:

1. We found many repetitive anchor stocks such as "APC" and "A" which affected more than one smaller stock in the "Top 100" report for any given day. This is not surprising since a strong stock may affect other stocks.

2. The most surprising finding is the fact that some dependent symbols such as "SHY" (**ISHARES LEHMAN 1**) were highly correlated to *multiple* larger capitalization stocks – "SHY" was a dependent stock in the 'Top 100' correlation runs 9903 times with 326 different anchor stocks, often more than once in a single day's "Top 100" report. Other similar dependent stocks were "AGG" and "BF.B".

What we can deduce from this second finding becomes more interesting since no logical "transitive" properties could be readily found – i.e. stock XYZ was dependent on larger capitalization stocks A, B and C in the same day, but no "Top 100" correlation could be found between any combination of A, B or C on that same day!

These logical anomalies are still under investigation….

Future Directions

Future phases could take several directions. Most definitely, we would expand our dataset to include all stocks from several disparate global markets. From there, we could limit the measurements and calculations for very specific changes:

1. Determine if higher correlations (or smaller error margins) can be found with smaller companies,
   a) against the indices
   b) against larger companies of the Nth degree.

2. Determine if any correlations can be found BETWEEN companies
   a) within the same market,
   b) within the same or related industries,
   c) explicitly across markets, and
   d) across otherwise non-related industries.

The amount of computing power needed to accommodate any future phase will be dependent on the number of stocks in the dataset as well as the limitations of measurement based on one or more of the above criteria.

We would welcome any collaboration requests to be able to continue this work.

Conclusions

I consider the outcome of this phase of the experiment to be a success – sufficient data patterns were found to support the algorithm works. The idea of crossing one mathematical culture with another will open up a whole new realm of research. The mathematics themselves are not original to each field, but the perspective in creating such algorithms and solving such solutions is a direct result of the culture from which they were born. Crossing cultures is what gives us new perspectives on old problems. Using mathematical theories such as polynomial adalines can open a new avenue for predictive analysis.

For Economics

Ultimately, this experiment was meant to prove (or conclusively disprove) that businesses of certain characteristics are more affected by the independent or seemingly unrelated global stock market activity than previously known economic factors.

The next step is to show that viewing the global market; *a company is by the decisions made in an unrelated industry*. With this knowledge, the company (or funds dependent on it) may be able to avoid vacillations in valuations by paying close attention to these "unrelated" correlations and being prepared.

Ideally, finding new correlations allows economists to look deeper into the factors that makeup this relationship and expand and refine new economic theories.

For Behavioral Biometrics

New purposes for the mathematics behind behavioral biometrics have been found – and the application of such algorithms can be tuned to the type of data they are meant to quantify. In this experiment, keystroke biometrics was used against the stock market because both are time sensitive. Would this be used to find special patterns in space? I would suspect handwriting recognition algorithms would make a better fit. Would this be useful to track extremely low-wave vibrations in the oceans? Again, voice recognition technology might be better suited for that task.

Conversely, can predictive analysis algorithms used with the stock market be transported for use in determining behavioral biometric factors – and provide a new source for unique individual identification?

Acknowledgements

We would like to thank Stephane Dubois at X:ignite for supplying market data in bulk, and customizing solutions to allow this experiment to go forward quickly and smoothly.

About the Author

John C. Checco, CISSP (john.checco@checco.com) is a member of the American Society for Industrial Security (ASIS) NYC Chapter and president of bioChec^{TM} (www.biochec.com), a division of Checco Services, Inc

Copies of the database results on CD can be obtained with written requests to:

StockWatch Request
Checco Services, Inc.
33 Capt Faldermeyer Drive
Stony Point, NY 10980-3463 USA