# Sliding Window Analysis of Binary *n*-Grams Relative Information for Financial Time Series

***Igor Borovikov*, Michael G. Sadovsky^***

**Center for Advanced Signal and Image Sciences (CASIS) at LLNL**

**18th Annual Workshop, May 21, 2014**

*) igor.borovikov@gmail.com (Nekkar.net: Int. Labs);
^) msad@icm.krasn.ru (Institute of Computational Modelling SB RAS)

# Abstract

The presentation describes a novel approach to statistical analysis of financial time series.

The approach is based on *n*-grams frequency dictionaries derived from the quantized market data. Here we focus on binary quantization. The frequency dictionaries are studied by evaluating their information capacity using relative entropy (Kullback-Leibler divergence). Calculating relative entropy in sliding window may lead to development of new market indicators, detecting market bubbles and other regimes.

Other possible applications of the proposed technique include market event study with the *n*-grams of extreme information value.

The finite length of the input data presents certain computational and theoretical challenges which we discuss.

This is a report on a work in progress and describes the methodology more than the final results.

# Introduction and Motivation

- markets considered not entirely random [1];
- ***technical analysis*** is viewed as a tool exposing non-randomness.
  - In its paradigm, asset price discounts all information available up until current moment to the market participants [3];
- analysis of ***information flow*** between market participants [2] may lead to new technical market indicators and insights into the market behavior;
- financial time series is the primary objects of this study and they thought to reflect the information flow in question;
- we use ***relative information*** methods inspired by novel works in bioinformatics [4,5,6] and build on our previous works [7,8].

Time series in this work represent adjusted close price *p(t)* where *t* is trading day. The series is obviously finite in both direction and real valued.

# Quantization Mapping

Transformations:

Adjusted Close $p(t) \rightarrow$ (log) Returns $r(t) \rightarrow \text{sign}^*(r(t))$

result in {-1,1}-valued series (binary quantization). The function *sign** here maps 0 to -1:

$$\text{sign}^*(x) = \begin{cases} 1 \text{ if } x > 0 \\ -1 \text{ if } x \leq 0 \end{cases}$$

Next, for convenience we use alphabet **A**={*a,A*} instead: -1→*a*, 1→*A*.

Thus, the series *p(t)* transforms into a string *s* in alphabet **A**, i.e. *Q*: *p(t)*→*s*.

Note: Other quantizations are possible. The described one is (a) the simplest one; (b) contains no parameters; (c) matches binary models in derivatives pricing.

# Sliding Window

We focus on substrings *w(t)* of fixed length ***W***<|s| that start at position t in the "master" string *s*. The quantization and substring ops commute.

# Dictionaries. Projection and Lifts.

**Frequency dictionary** $D_n(w)$ for string $w$ is mapping:

$D_n: v \to f_v,$

where:

$v$ is an $n$-gram $v$ of length $n$ in alphabet $\mathbf{A}$,

$f_v$ is its normalized frequency in the string $w$ (i.e. $\sum f_v = 1$).

We call $n$ in this definition **dictionary thickness.**

**Projection** $\mathcal{P}_k: D_n \to D'_{n-k}$ of dictionary $D_n$ to dictionary $D'_{n-k}$ is natural mapping that computes frequencies of $(n-k)$-grams from $D_n(w)$ rather than from the original string $w$.

**Lift** $\mathcal{L}_k: D_n \to D^*_{n+k}$ of dictionary $D_n$ is *right* inverse mapping:

$\mathcal{P}_k \circ \mathcal{L}_k = 1.$

Notes: (a) Looped and unlooped strings $w$ lead to subtly different definitions for lifts and projections. We ignore that here. (b) Lift is not defined uniquely, which is easy to see. (c) Generally $\mathcal{P}_k \circ \mathcal{L}_k \neq 1$.

# Max Entropy Lift

**Dictionary Entropy**: $S(D_n) = -\sum f_v \ln f_v$

**Max Entropy Lift** $\mathcal{L}^{\circ}_{k}(D_n)$: $D^{\circ}_{n+k} \dot{=} \text{argmax } S(D^*_{n+k})$ s.t. $\mathcal{P}_k(D^*_{n+k}) = D_n$.

Max Entropy Lift $\mathcal{L}^{\circ}_{k}$ is unique and can be computed using Lagrange multipliers where constraints are linear equations on frequencies of n-grams in $D^*_{n+k}$ and $D_n$ [4,5,6].

# Information Capacity

Denote $\tilde{D}_n = \mathcal{L}^{\circ}_{1}(D_{n-1})$. (=max entropy lift from the (n-1)-grams dictionary)

**Information capacity** $S_n$ of string s on n-grams of length n is relative entropy of $D_n$ against $\tilde{D}_n$ (Kullback-Leibler divergence):

$$S_n = \sum f_v \ln (f_v / \tilde{f}_v),$$

where $f_v$ and $\tilde{f}_v$ are frequencies from $D_n$ and $\tilde{D}_n$.

# Normalized Information Capacity

If $S'_n$ is bootstrapped (via an equivalent Bernoulli process) information capacity then normalized information capacity $\overline{S_n}$ is defined in terms of expectation E and standard deviation $\sigma$ of the bootstrapped value:

$$\overline{S_n} = \frac{S_n - E(S'_n)}{\sigma(S'_n)}$$

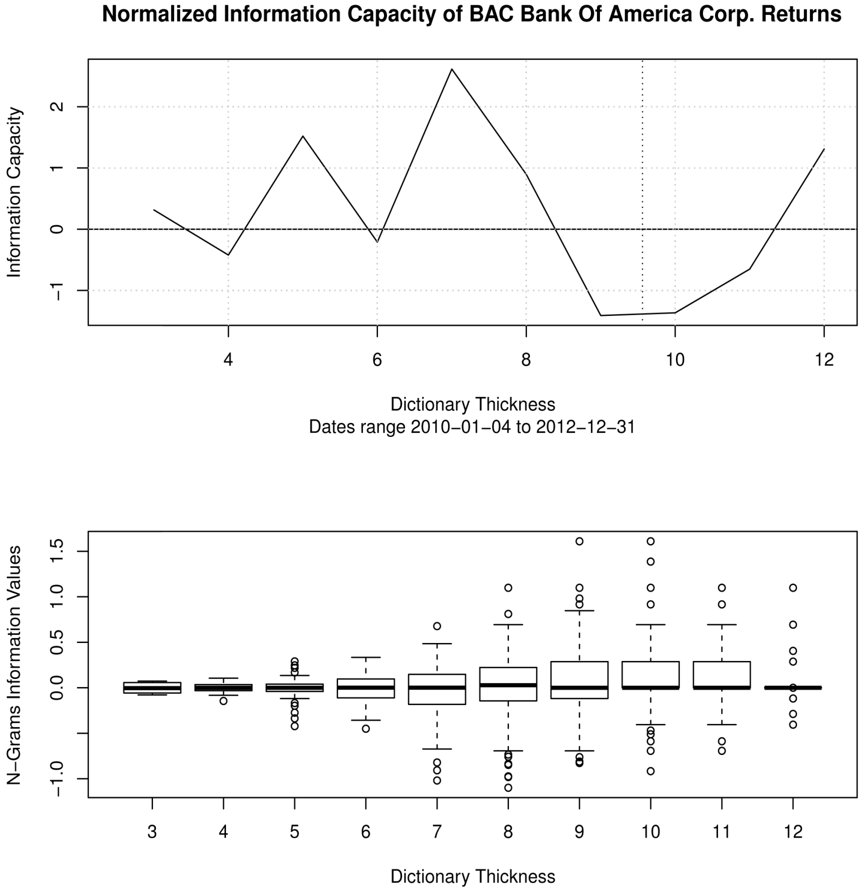

**Figure 1.**
Top: BAC (Bank of America) normalized information capacity computed for single day return, static window W=500.
Vertical dashed line shows noise limit (see below).
Bottom: box plot of normalized inf value of n-grams.

# “Noise Limit” from Finite Window Length

There are $\boldsymbol{W} - n$ of n-grams in the text of length $\boldsymbol{W}$ and $2^n$ possible $n$-grams in alphabet $|\mathbf{A}|=2$. Hence we keep $n < \log_2 \boldsymbol{W}$.

> Note: More advanced analysis using autocorrelation function of $n$-grams can give better estimates.

# n-Grams with Extreme Information Value

If $f_v$ and $\widetilde{f_v}$ are frequencies ($\neq 0$) of n-gram $v$ in $D_n$ and $\widetilde{D_n}$, then:

- $v$ is ($\varepsilon$-)**information rich** if $|\ln (f_v / \widetilde{f_v})| \geq 1/\varepsilon$,
  (IR n-gram for brevity),
- $v$ is ($\varepsilon$-)**information poor** if $|\ln (f_v / \widetilde{f_v})| \leq \varepsilon$,
  (IP n-gram for brevity)

for some $1 \geq \varepsilon > 0$.

Hypothesis: Both IR and IP n-grams may indicate specific states of the market associated with trends, their beginning and end.

# Multiple Days Returns

Single day returns are subject to noise. We also consider multiple days $d$ returns while keeping $d$ (much) smaller than $\boldsymbol{W}$ (window size).

# Aggregate Values Calculation

For a fixed sliding window position, summation of information value, head n-gram count and percentile over $n$ in range $2<n<$noise limit and return days in range $d < k\boldsymbol{W}$ can be used to reduce the noise in data.

In addition to information value we calculate:

- Head n-gram count is number of occurrences of the head n-gram inside the current window;
- Head percentile is percentage of the n-grams which have higher information value than the head n-gram.

Note: $k$ is set to approximately 0.2 to keep noise limit about the same.

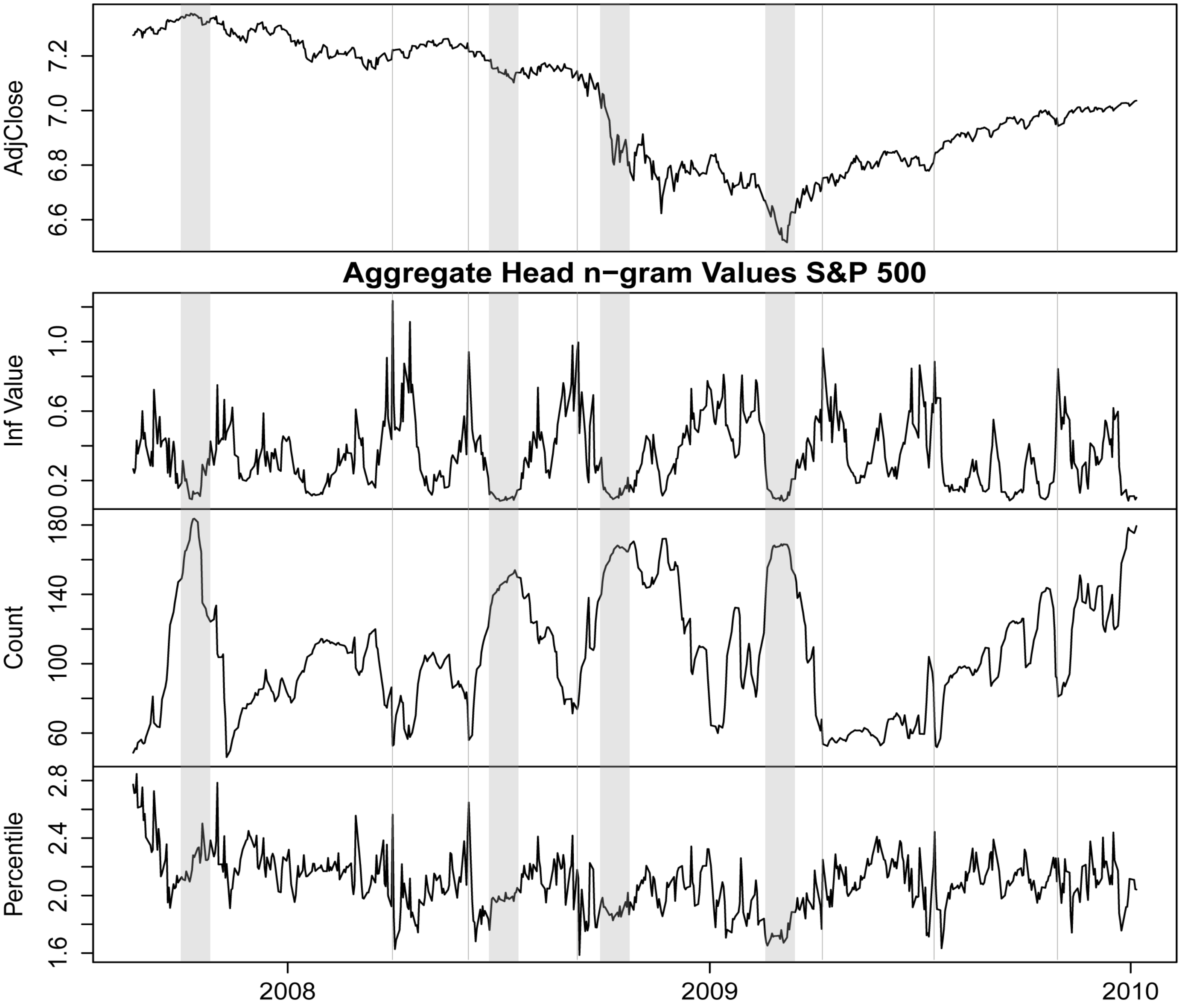

**Figure 2.** Aggregate information value, count and percentile. S&P Crash of 2008. See legend on page 11.

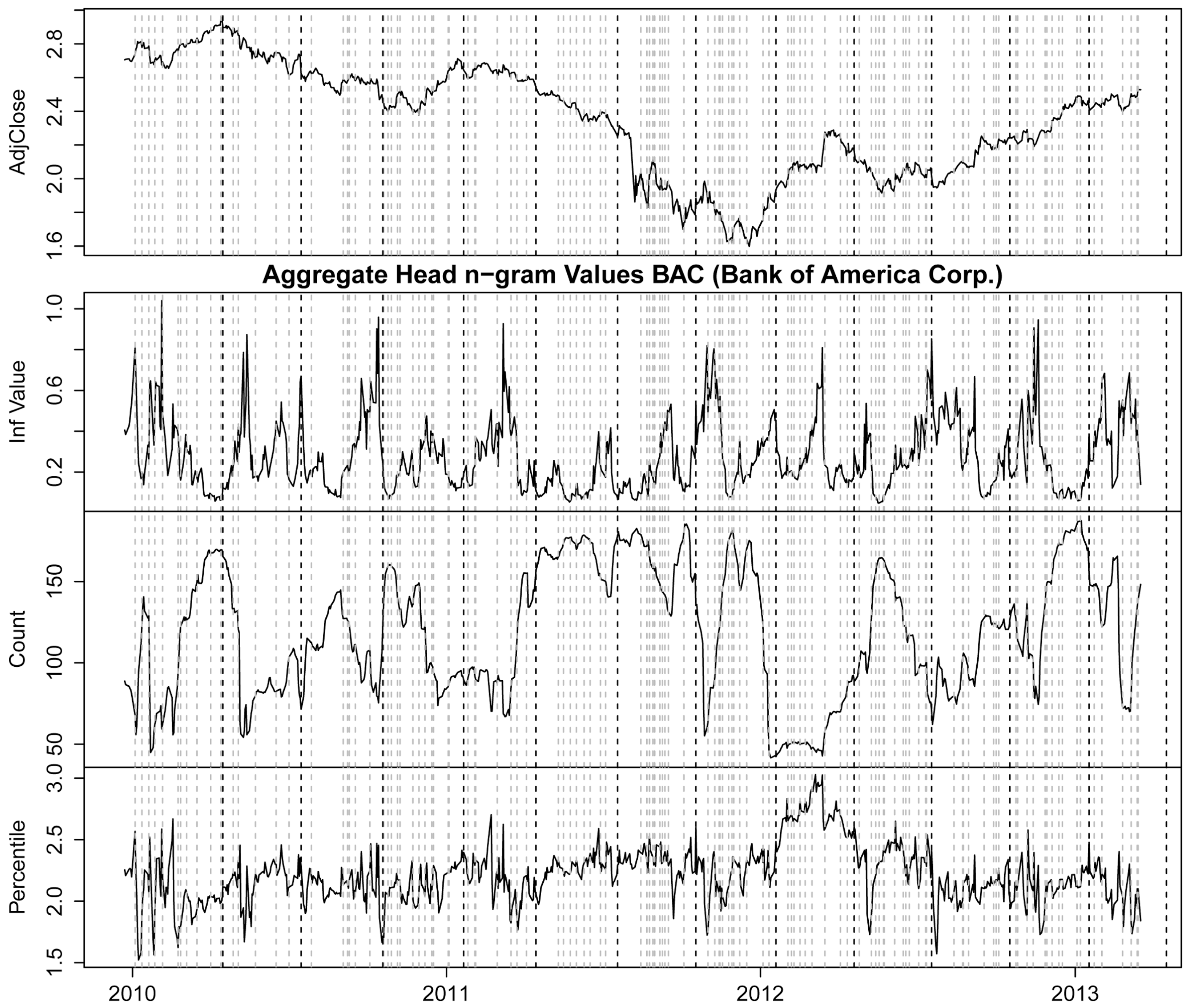

**Figure 3.** Aggregate information value, count and percentile: official announcements. See legend on page 11.

Figure 2 legend: S&P 500 around the crash of 2008. *Vertical Lines*: notable spikes (IR n-gram events) appear to show in the beginning of a new trend? *Vertical Bars*: troughs in head information value (clustered IP n-grams) possibly indicate end of a trend?

Information value and Count are strongly anti-correlated.

Figure 3 legend. BAC official announcements overlayed on aggregate values: quarterly reports as black vertical lines, other announcements as vertical grey dashed lines.

# Discussion

While Figure 2 may suggest connection between notable trends and appearance of IR and clustered IP *n*-grams, the result is inconclusive and requires a formal definition of a trend and proper statistical verification.

Figure 3 fails to show any obvious connection between aggregate information value and announced events. This may be due to many reasons: lack of surprize in some of the events or their insignificance to the investors.

Future work will focus on making this findings more concrete.

# References

[1] A.W.Lo, A.C.Mc Kinlay, A Non-Random Walk down Wall Street, Princeton Univ. Press, 1999.

[2] E.F.Fama, L.Fisher, M.C.Jensen, R.Roll, The adjustment of stock prices to new information, Int. Economic Review, 10(1969), no.1, 1–21.

[3] J.J.Murphy, Technical analysis of the financial markets. A comprehensive guide to trading methods and applications, New-York institute of finance, 1999.

[4] N.N.Bugaenko, A.N.Gorban, M.G.Sadovsky, Towards the definition of information content of nucleotide sequences, Molecular biology Moscow, 30(1996), no. 5, 529–541.

[5] N.N.Bugaenko, A.N.Gorban, M.G.Sadovsky, The information capacity of nucleotide sequences and their fragments, Biophysics, 5(1997), 1063–1069.

[6] N.N.Bugaenko, A.N.Gorban, M.G.Sadovsky, Maximum entropy method in analysis of genetic text and measurement of its information content, Open Systems & Information Dyn, 5(1998), no, 2, 265–278.

[7] Igor Borovikov, Michael G. Sadovsky, Analysis of financial time series with binary *n*-grams frequency dictionaries, *J. Sib. Fed. Univ. Math. Phys.*, 2014, **7**:1, 112–123

[8] Igor Borovikov, Michael Sadovsky, A relative information approach to financial time series analysis using binary *n*-grams dictionaries, arXiv:1308.2732, 2013, [q-fin.ST] 13 pages.