Journal of Siberian Federal University. Mathematics & Physics 2014, 7(1), 112–123

УДК 51:336 + 330.47

Analysis of Financial Time Series with Binary N-Grams Frequency Dictionaries

Michael G. Sadovsky*
Institute of computational modelling SB RAS,
Akademgorodok, Krasnoyarsk, 660036 Russia

Igor Borovikov†
Nekkar.Net Labs, Ltd.
California, USA

Received 10.06.2013, received in revised form 10.08.2013, accepted 05.09.2013

The paper presents 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. Such dictionaries are studied by evaluating their information capacity using relative entropy. A specific quantization of (originally continuous) financial data is considered: so called binary quantization. Possible applications of the proposed technique include market event study with the n-grams of higher information value. The finite length of the input data presents certain computational and theoretical challenges discussed in the paper. also, some other versions of a quantization are discussed.

Keywords: order, entropy, mutual entropy, indicator, trend.

Introduction

A key idea of any research targeted to financial data mining is to figure out some order in them, and elabourate a number of indicators to predict some (important) events at the stock markets. Mathematical modelling and relevant mathematical methods were widely applied in this area. Currently, this field of studies is knows as *technical analysis* and has a good history behind with an effective culture of doing research here (see, e. g, [1, 2]).

Money flow, with neither respect to its specific form, could be considered as a time series, either discrete, or continuous, and relevant mathematical techniques of the analysis could be implemented. A good starting reading could be found in [2,3]. Finally, the classics by J. Murphy [4] must be cited as a key reading for anyone interested in the foundations of the subject‡. Basically, various techniques [5] and approaches of (linear) statistics analysis and probability theory [6] are implemented. Empirical studies [7] are also important.

In this paper we follow the approach initially developed for the analysis of genetic texts in the pioneering works [8, 9]; this approach seems to be novel, for financial series analysis. The main idea is to build n-gram frequency dictionaries $D(n)$ from a sufficiently large input text(s) for n-grams of a different length n. The entropy-maximization procedures described in the cited works yields the new dictionaries $D^k(n+k)$ from $D(n)$, $k = 1, 2, \ldots$. These are called reconstructed (or lifted) dictionaries. The Kullback-Leibler divergence between reconstructed $D^k(n)$ and the original $D(n)$ dictionaries for the same n gives a *relative* information capacity of the input text§

*msad@icm.krasn.ru
†igor.borovikov@gmail.com

‡There are the translations into Russian of his books.
§Actually, the information capacity is definded for a frequency dictionary rather than for a text; further we shall not distinguish this point.

for the n-grams of length n. Such analysis also results in the detection of "divergent" n-grams responsible for "higher information content". The definition of information capacity introduced this way is not equivalent to the Shannon's or Boltzmann's classical ones, which are based on *absolute* entropy of the text. It is worth noting that the n-grams frequency dictionary approach is not using any explicit assumptions regarding the text, like Markovian property, or alphabet letters distribution.

Next, following the cited works, we explain information capacity defined with the n-grams frequency dictionaries. This leads to selecting the optimal length l_{opt} of n-grams for further analysis as the length yielding the maximum of (normalized) information content over the all n-gram lengths and possibly other parameters of the method. The intuition behind this is that for the optimal set of parameters, the l_{opt}-grams are the least predictable ones. We attempt to connect them to significant market events and/or trends. The normalization of the information content turns out to be more important task comparing to bioinformatics since financial time series usually result in much shorter input texts than a typical sequenced genome.

1. N-grams dictionaries from time series

To avoid an ambiguity, we shall use the term *ticker* when talking about a security like company shares, ETFs or indices (e.g. GOOG, YANG or ^DJI). We will reserve the terms *letter* and *symbols* (to be used interchangeably) for the elements of the alphabet we are going to construct.

We consider the simplest case of a financial time series, namely Adjusted Close daily price on a ticker[¶] denoted by $z(t)$, from which we calculate either log- or simple returns $p(t)$:

$$p(t) = \log(z(t)/z(t-1)) \approx \frac{z(t)}{z(t-1)} - 1 \ .$$

Here Adjusted Close price $z(t)$ is a real number and t is (trading) day treated as an integer index. The choice of simple returns over log-returns is not critical for this work so we will not distinguish them further.

1.1. Constructing texts from time series

To apply the n-grams-based methods, we have to represent time series $p = p(t)$ as a (very long) word in an alphabet $\aleph$. We shall call this word *input text*. The letters of the alphabet encode quantized values of $p(t)$. The choice of mapping $R \to \aleph$ (here $p(t) \in R$) and the choice of the alphabet $\aleph$ are the parameters of method. This paper is concentrated on the special case of the mapping into a binary alphabet $\aleph = \{0, 1\}$. Some other types of mappings will discussed later (see Section 3.). In particular, we focus on alphabets with even number of letters corresponding to different multiples of up-ticks and down-ticks in the price movement.

Definition 1. *A finite alphabet $\aleph_N$ of the cardinality $2N > 0$ is called an* output alphabet *if it is ordered by bijective mapping to the set of integers $Z_N = \{-N, -(N-1), \ldots, -2, -1, 1, 2, \ldots, N-1, N\}$ (note the absence of 0). The mapping $X\colon Z_N \to \aleph_N$ is called* indexing.

Binary dictionary is a special case: it can be represented with the digits $\{0, 1\}$ where 0 corresponds to the negative values of p and 1 corresponds to the positive values. Note that quantization includes clamping of input values. In the most straightforward approach, this results in mapping of the entire ranges $(-\infty, P_{-N}]$ and $[P_m, +\infty)$ to the first and the last letters of the alphabet, correspondingly. This is useful when we need to limit the alphabet cardinality

[¶]The source of the data used throughout this work is the publicly available financial data from *Yahoo!*Finance unless indicated otherwise.

introducing no additional complicated non-linearity into quantization mapping. The binary alphabet is a trivial example of clamping with $N = 1$ and $P_{-1} = P_1 = 0$.

To summarize, the base procedure that generates an input text from a series $z(t)$ of Adjusted Close prices consists of the three steps:

1) Convert prices time series $z(t)$ to (log-) returns $p(t)$,
2) Specify the output alphabet $\aleph$ and the quantization mapping Q: $R \to \aleph$,
3) Quantize $p(t)$ to obtain the text $\mathsf{T} = \{Q(p)\}$.

1.2. Dictionaries from the input text

Given an input text T of a finite length L, first we build natural frequency dictionary $D(n)$ by counting all n-grams occurrences C_w for each n-gram w in the text T. It yields a set of pairs (w, C_w). Let C_* be the total number of n-grams in T. Obviously, $C_* = |\mathsf{T}| - n$, where $|\mathsf{T}|$ is the text length L. Normalization by C_* gives the *frequency* of the n-gram w: $f_w = C_w/C_*$.

Definition 2. *The (natural)* frequency dictionary $D(n)$ *of the text* T *is the set of all pairs* $\{(w, f_w)\}$ *where* w *are unique* n*-grams and* f_w *are the corresponding frequencies constructed as described above. The parameter* n *is called the* thickness *of the dictionary. The set* $\Omega = \{w\}$ *is called the* support *of the dictionary.*

It should be said that any text T could be unambiguously converted into a frequency dictionary; an inverse does not hold true, in general. Indeed, a set $S(l) \ni w$ of strings (of the given length l) assigned with the positive real numbers f_w so that $\sum_{w \in S} f_w = 1$ may correspond to a neither text. Since we have no aim to address a problem of a reconstruction of entire text from a dictionary, we shall not consider this issue, further.

A dictionary $D(n)$ of the thickness n can be naturally projected into the frequency dictionary $D_1(n)$ of thickness $n-1$ bearing $(n-1)$-grams and their (reciprocal) frequencies. More generally, we can compute the dictionary $D_k(n)$ bearing $(n-k)$-grams with the reciprocal frequencies. It is a straightforward procedure that calculates all $(n-k)$-grams and their frequencies not from the original text T but rather from $D(n)$ with proper counting of the corresponding frequencies. This procedure uniquely defines the operator $\mathbb{P}_k$: $D(n) \to D_k(n)$ for k ranging in $0, \ldots, (n-1)$; here $\mathbb{P}_0$: $D(n) \equiv D(n)$. Further, such downward transformation will be denoted with a lower index; thus, $D_k(n) = \overline{D}(n-k)$ is indeed a frequency dictionary of the thickness $n-k$.

The inverse upward operator $\mathbb{L}_k$: $D(n) \to D^k(n)$ makes a frequency dictionary $D^k(n) = \overline{D}(n+k)$ of the thickness $n+k$ from the dictionary $D(n)$; here $k > 0$ is an arbitrary positive integer. One can easily see that $D^k(n)$ is not uniquely defined; a family of different dictionaries $\{\overline{D}(n+k)\}$, instead. Any dictionary from the family yields the original frequency dictionary $D(n)$ due to an operator $\mathbb{P}_k$ execution: $\mathbb{P}_k[\overline{D}(n+k)] \to D(n)$, $\forall \overline{D}(n+k) \in \{\overline{D}(n+k)\}$. In such capacity, the operators $\mathbb{P}_k$ and $\mathbb{L}_k$ are not commutative ones:

$$(\mathbb{P}_k \circ \mathbb{L}_k) : D(n) \to D(n) ; \qquad (\mathbb{L}_k \circ \mathbb{P}_k) : D(n) \to ? .$$

To address this problem, one has to choose some peculiar frequency dictionary $\widetilde{D}(n+k)$ from the family $\{\overline{D}(n+k)\}$ of the extended ones. It should be kept in mind, that the family $\{\overline{D}(n+k)\}$ consists of various frequency dictionaries $\overline{D}(n+k)$, and the natural one $D(n+k)$ is among them. Here the maximum entropy principle may bring a solution, see [8–10] for the details and proofs. A brief outline of these results follows in the subsection 1.

It turns out that the comparison of the same-thickness dictionaries $D(n)$ and $D^k(n-k)$ (i. e. extended dictionary vs. the natural one) provides the grounds for useful insights into statistical properties of the text T, which are not readily accessible by other means.

Note that the original works [8–10] considered circularly looped input texts for the dictionaries generation. Here we can not require any periodicity of the input text T because it will create

artificial connection between otherwise disconnected trading days at the beginning and at the end of the analyzed time interval. The absence of the loop will create a complication to be discussed later but for now we will just ignore it. The approximation by the results from the looped texts improves as $|\mathsf{T}| \to \infty$.

1.3. Reconstructed dictionary and the information valued n-grams

Again consider an input text T defined over a finite alphabet $\aleph$. We can construct a sequence of dictionaries $D(j)$ of increasing thickness j:

$$D(1) \leftrightarrow D(2) \leftrightarrow \ldots \leftrightarrow D(j) \leftrightarrow D(j+1) \leftrightarrow \ldots \leftrightarrow D(L)\,. \tag{1}$$

The projection operator $\mathbb{P}_k$ (arrows pointing left in (1)), i. e., the transition $D(j) \mapsto D(j-1)$ is unambiguous. The opposite operator (that is $\mathbb{L}_k$) is ambiguous, generally, since an n-gram w may have multiple valid continuations (not more than the cardinality of $|\aleph|$).

A *valid* 1-*lift* is a transformation L_1: $D(j) \mapsto W(j+1)$ so that $W(j+1)$ is a dictionary of thickness $n+1$ and P_1: $W(j+1) \to D(j)$. So, by definition, a valid 1-lift L_1 satisfies $P_1 \circ L_1 = I$, where I is the identity mapping of $D(j)$. Thus, a lifted (extended) dictionary consists of n-grams $w \in \Omega$ extended by adding a prefix or a suffix of length 1 in the way that the projection of that former yields the original frequency dictionary $D(n)$. Obviously, adding an infix to the original n-grams one may not get a valid lift.

In other words, each combined set $f^*_{\nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q\nu_{q+1}}$ of the extended n-grams must satisfy the constraint

$$\sum_{\nu_{q+1}} f^*_{\nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q\nu_{q+1}} = \sum_{\nu_{q+1}} f^*_{\nu_{q+1}\nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q} = f_{\nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q}\ , \tag{2}$$

where $f_{\nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q}$ is the frequency of an n-gram $w \in D(q)$ in the original frequency dictionary $D(q)$. Linear constraints (2) eliminate some of the possible extensions for the original n-grams, but still do not define the lift uniquely.

As the final step to define the lift uniquely we shall use the maximum entropy principle:

$$\max_j \left\{ -\sum_{w^*} f^{(j)}_{w^*} \ln f^{(j)}_{w^*} \right\}. \tag{3}$$

Here $w^* = \nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q\nu_{q+1}$ denotes an n-gram satisfying the linear constraint (2), and j enlists the versions of feasible extensions. The maximum-entropy dictionary $\widetilde{D}(q+1)$ satisfying both (2) and (3) exists always, since the set of the dictionaries to be constructed from the given one is finite.

The frequency of the n-grams in the max-entropy lift $\widetilde{w} \in \widetilde{D}(q+1)$ could be computed explicitly using La Grange multipliers method [8–10]. It is determined by the expression

$$\widetilde{f}_{\nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q\nu_{q+1}} = \frac{f_{\nu_1\nu_2\nu_3\ldots\nu_{q-1}\nu_q} f_{\nu_2\nu_3\ldots\nu_{q-1}\nu_q\nu_{q+1}}}{f_{\nu_2\nu_3\ldots\nu_{q-1}\nu_q}}\ . \tag{4}$$

Similarly, the maximum entropy principle (3) allows to reconstruct the dictionary $\widetilde{D}(q+l)$ for l-lifts of any $l > 1$, see [8–10] for details.

The 1-lift to a thicker dictionary via (4) yields the dictionary that bears no "additional" information, external with respect to that one contained in the original dictionary. It consists of the n-grams of the length $q+1$ that are the most probable continuations of the strings of the length q. The lifted dictionary $\widetilde{D}(q+1)$ contains all the strings that occur in the original

dictionary $D(q+1)$ and, possibly, some other ones. For any q, $q \geqslant 1$ the following inequality of the entropy:

$$S\left[\widetilde{D}_{q+1}\right] \geqslant S\left[D_{q+1}\right]$$

holds true.

The maximum entropy approach generalizes to valid l-lifts, $l > 1$, and also results in a unique solution. Everywhere below we shall focus on 1-lifts, only; besides, no other techniques of the lifting would be considered, but the max-entropy lift.

1.4. Information capacity of a text

Here we outline the idea of the information valuable n-grams (see sec. 1.). Consider two sequences of the frequency dictionaries: the one of the dictionaries constructed directly from the input text (the natural dictionaries)

$$D(1) \leftrightarrow D(2) \leftrightarrow \ldots \leftrightarrow D(j) \leftrightarrow D(j+1) \leftrightarrow \ldots \leftrightarrow D(L)$$

and the other sequence

$$\widetilde{D}(2) \leftrightarrow \widetilde{D}(3) \leftrightarrow \ldots \leftrightarrow \widetilde{D}(j) \leftrightarrow \widetilde{D}(j+1) \leftrightarrow \ldots \leftrightarrow \widetilde{D}(L)$$

of the lifted dictionaries. Here we assume that $\widetilde{D}(j)$ is always lifted from $D(j-1)$, $j = 2, \ldots, L$.

Definition 3. Information capacity $\overline{S}_j$ *of a natural dictionary* $D(j)$ *is the mutual entropy*

$$\overline{S}_j = \sum_{w \in \Omega} f_w \ln\left(\frac{f_w}{\widetilde{f}_w}\right) \tag{5}$$

of the natural dictionary $D(j)$ *calculated against its lifted-up entity* $\widetilde{D}(j)$ *derived from the dictionary* $D(j-1)$.

This definition is applicable to any valid lifts. For the case of (4) (max-entropy lift), the information capacity could be easily determined:

$$\overline{S}_j = 2S_{j-1} - S_j - S_{j-2} \qquad \text{and} \qquad \overline{S}_2 = 2S_1 - S_2\,, \tag{6}$$

where S_j is absolute entropy of the natural dictionary $D(j)$.

1.5. Information valuable (divergent) n-grams

Consider again the information capacity (5). Sufficiently close values of natural frequencies f_w and lifted frequencies $\widetilde{f}_w$ of the same n-gram w make smaller contribution (per n-gram) to the overall value of the sum, while the n-grams with the greatest deviation provide greater-than-average contribution. This observation motivates the following:

Definition 4. Information valuable n-gram $\widehat{w}$ *(an element of the frequency dictionary* D_j*) is an* n*-gram satisfying*

$$|\log f_{\widehat{w}} - \log \widetilde{f}_{\widehat{w}}| > \log \alpha\,,$$

where $1 \geqslant \alpha > 0$ *is the information value threshold.*

We will also call such n*-grams* α-divergent n*-grams, or* divergent n*-grams when parameter* α *is obvious from the context or its specific value is not important.*

The complement to the subset of the divergent n-grams is the subset of α-*ordinary* n-grams (or just *ordinary* n-grams). If $\alpha = 1$ then all the n-grams within the dictionary D_j are divergent with the exception of those whose lifted frequency exactly equal the natural one. Usually such n-grams occur only in very long input texts.

The definition 4 includes the parameter α and its practical choice depends on application. Making α large enough, so that the count C_w of at least some of the divergent n-grams w found in the dictionary is greater than 1, provides a reasonable guideline to setting minimal practical threshold. Such choice ensures that found divergent n-grams are not all degenerate (i. e. produced by combination of short input text and large n). It was said above such unique degenerate n-grams normally should be excluded from the analysis.

There are several issues stemming from the finiteness of the length of an input text. Since we shall use the divergent words as a tool in our studies of the financial time series, we should pay more attention on their features and obstacles arisen from the finite sampling, and discuss them immediately.

1.5.1. Noise barrier

The first issue we have to deal with is related to the noise resulted from the finite length of an input text to affect the figures of information capacity. We have already touched this subject briefly earlier but a consideration of some more details would be helpful for further analysis of financial data (see Section 2.).

To simplify the issue, consider a binary alphabet with proximal probabilities of both symbols. The total number of *different* n-grams of length n is then 2^n. There are total $L - n + 1$ of *all* n-grams of the length n in the input text of the length L. For $L \gg n$ we can take the number of n-grams $\approx L$. If $L = 2^n$ then each occurrence of each n-gram is "critical" in a sense that every difference of lifted dictionary from the natural one will "amplify" the random nature of the input text.

When dictionary thickness exceeds $\log_2 L$, some of the n-grams will not be present at all (go "extinct"). This follows in a degeneration of the information capacity as many of the longer n-grams will be lifted from the shorter ones uniquely and other will not be present. The number of uniquely reconstructed n-grams grows up as the ratio L/n gets smaller. This results in the bell curve of information capacity discussed earlier. Its peak is located at the value of n close to $j_{\max} \approx \log_2 L$.

Similarly for the alphabets of cardinality k the figure $j_{\max} \approx \log_k L$ approximates the location of the peak. Again, this approximation holds better when for the probability density of the letters in the alphabet are close to the uniform and the value $j_{\max}$ gets smaller when the distribution is far from the uniform. We will call the figure of $j_{\max}$ *the noise barrier* for the given input text length.

1.5.2. Normalized information capacity

To separate the actual signal in the information capacity $\overline{S}_n$ of the input text T from the noise, we need to compare it to the expectation $\mathsf{E}(S'_n)$ of the information capacity S'_n calculated from the randomly generated surrogate texts T'. Random texts T' must have the same length and yield the same probabilities of the alphabet letters as T does. The figure of an absolute difference of the values S_n and S'_n are not as useful, as the normalized one by the standard deviation $\sigma(S'_n)$ of the information capacity of the corresponding random input texts T'. This normalization gives σ-distance from the purely random signal.

Definition 5. *The* normalized information capacity *of the input text* T *is defined as*

$$S^*_n = \frac{\overline{S}_n - \mathsf{E}(S'_n)}{\sigma(S'_n)}\,, \tag{7}$$

where $\overline{S}_n$ *is the information capacity of the original input text;* $\mathsf{E}(S'_n)$ *and* $\sigma(S'_n)$ *are the expectation and the standard deviation correspondingly of the information capacity of the random input text* T' *so that* $\mathsf{E}(D(1)(\mathsf{T}')) = D(1)(\mathsf{T})$*. The last condition means that the the letters in the source of the random texts* T' *are distributed in the same way as for the original text* T*.*

In practice, to estimate the values of $\mathsf{E}(S'_n)$ and $\sigma(S'_n)$ one should go through the following steps using Monte-Carlo method:

1) Compute probability distribution $D(1)$ of the letters in the input text;
2) Generate a set $\{\mathsf{T}'_k\}$ $k = 1, \ldots, M$ of sufficiently large number M of the random texts T' of the same length generated using probabilities $D(1)$;
3) For T' estimate $\mathsf{E}(S'_n)$ and $\sigma(S'_n)$.
4) Calculate the normalized value (7).

The complexity of this method is obviously exponential. That makes the parameter $|\mathsf{T}| = N$ that is the length of an input text important one in terms of the computational costs.

2. Case studies, aliasing and the relation to ACF

We start the analysis with a broad index Russel 2000 that may represent to certain extent a typical behaviour of broad market returns. Then we move to the behaviour of historical returns of Bank of America (ticker BAC) while touching on some other tickers. In further discussion all the input texts have been generated via binary quantization. The typical time window for the analysis was three years.

2.1. Historical Returns of Russel 2000 (^RUT); ACF vs. Information Capacity

Figure 1 shows log of the absolute values of the information capacity calculated for the ticker ^RUT. We used logarithmic scale for the plot since the information capacity has different order of magnitudes at the different dictionary thickness figures. There were approximately 750 trading dates worth of data in the represented time window. It gives the noise limit value between 9 and 10. The peak of information capacity is observed close to the dictionary thickness 10, as expected.

There are few more important observations coming from Fig. 1. The first observation indicates that the behaviour of the wide market returns is quite close to the noise and does not exhibit any dramatic deviations as the solid line lies within single σ from the noise signal.

The second observation is that auto-correlation (ACF) and the information capacity plotted together show no obvious connection. This suggests that n-grams-based analysis is not directly tied with ACF-based one and represents a new statistical aspect of the input text. Yet there must be some connection between information capacity and ACF. The following simple argument supports this hypothesis. Consider periodic input text T with the period L and the length $|T| \gg L$. Obviously, the information capacity of such periodic text degenerates (becomes 0) on the dictionary thickness $> L$. And the ACF has maximum at the value L. The authors will revisit the connection between information capacity and auto-correlation in the future works.

A quick comparison of log-information capacity (Fig. 1) and normalized information capacity on the Fig. 2 shows the advantages of the normalized representation for the information capacity. Additionally, the Fig. 2 shows box-plot for the n-grams in the different thickness dictionaries. It is interesting to visually confirm that the n-grams distribution degenerates into multiple outliers and the condensed central peak for the values above noise limit (vertical dotted line on Fig. 2). The overall explorative analysis using information capacity for the ticker ^RUT indicates quite good correspondence with the random market hypothesis.

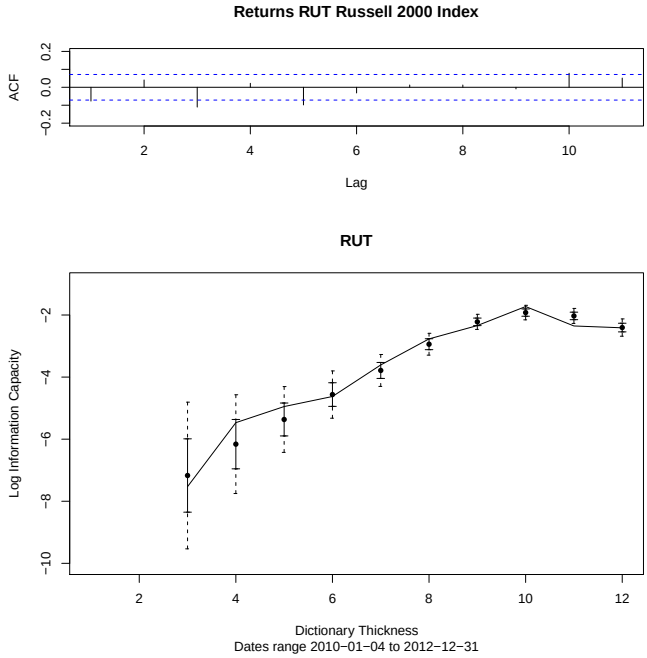

Fig. 1. Upper chart: auto-correlation (ACF) of daily log-returns. Lower chart: solid line shows *log*-information capacity, dots show $\mathsf{E}(S'_n)$ (the expectation of the information capacity), solid vertical bars mark $\sigma(S'_n)$ and dotted vertical bars mark $2\sigma(S'_n)$ as in Definition 5

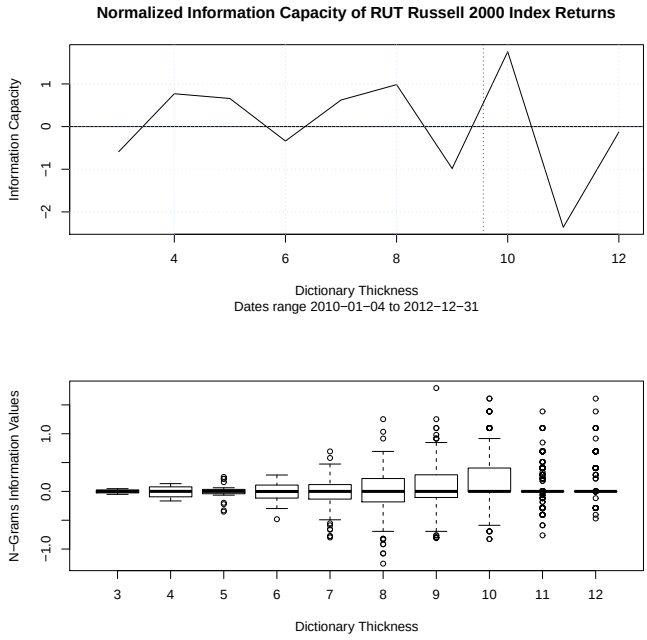

Fig. 2. Upper chart: normalized information capacity. Lower chart: box plot for the n-grams distribution of the input text. The vertical dotted line indicates the noise limit of the input text

2.2. Case studies: Starbucks Corp. (ticker SBUX) and Bank of America Corp. (ticker BAC)

Starbucks Corp. shows more interesting example, which is less in line with the hypothesis of random behaviour of returns. This manifests in high deviation from the information capacity of the corresponding noise and somewhat slower degeneration of the n-grams distribution for the values n close to the noise limit. This last observation is supported by the box-plots. Another example is provided by Bank of America Corp. (ticker BAC). Its information capacity exhibits significant deviation from the average information capacity for $n = 7$.

The less noisy behaviour of the tickers for particular companies was discovered in many more cases. This hardly is unexpected. The behaviour of the returns for particular companies is heavily influenced by the events significant for the company (e.g. earning reports, other announcements from the company and the competition, etc). As such, particular company returns are less prone to the averaging to the pure noise, unlike the wide market returns where such events are more spread out in time and diluted by the large number of the index components.

2.3. Aliasing

Even though the signal for SBUX is unusually strong at $n = 3$, we can't attribute it entirely to some source of meaningful information. The reason for that comes from the very limited cardinality of the dictionary D_3. For the binary alphabet $|D_3| = 2^3 = 8$, i. e. we have only

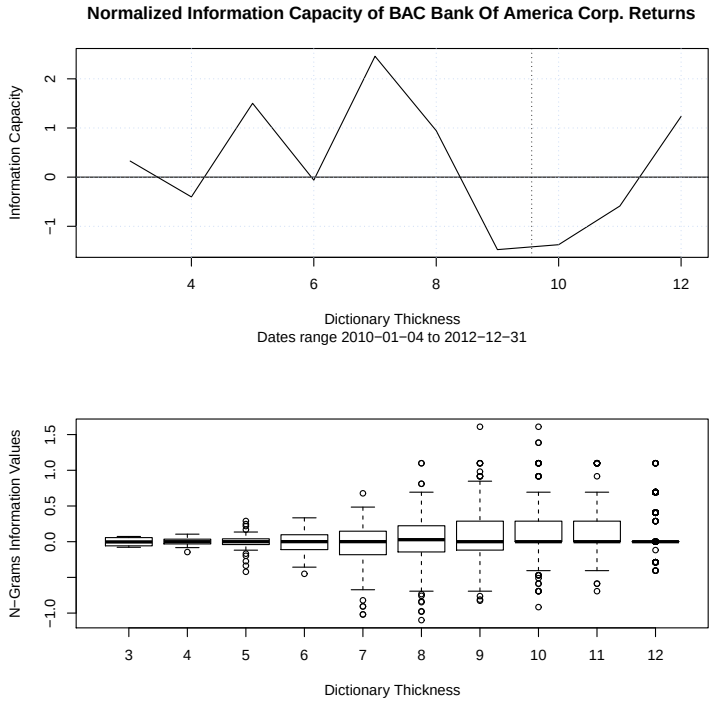

Fig. 3. Ticker BAC (Bank of America Corp.): the information capacity is outside of 2σ distance from the corresponding noise for $n = 7$

8 different n-grams available for the analysis of information capacity. It translates into the following: all of the sequences of the length 3 contained in the original input series of the returns

are mapped via quantization mapping to only 8 available n-grams. Thus truly significant triplets of the market days for the ticker get binned together with the ordinary ones only due to the lack of the unique n-grams in the target quantization space. This effect is similar to the aliasing in image and signal processing when limited target quantization space generates unwanted artifacts and leads to the additional information loss.

Such aliasing also makes the choice of the α-divergent n-grams for small values of n difficult or impossible because the meaningful choice of the threshold value α (see definition 4) should result in non-unique n-grams. BAC shows strong aliasing at $n = 3$: the information capacity of n-grams is batched into two clusters on the extremes of the range leaving a tighten choice for α figures in α-divergence.

2.4. Divergent n-grams: distribution over time

The location of divergent n-grams on the time axis can be studied from the point of view of their possible connection with the market events or particular price action of the underlying ticker. The divergent n-grams were picked in such way that none of them was unique and they represent no less than 5% of the total number of the n-grams. For small n that resulted in higher percentage due to the aliasing discussed earlier in sub-section 2..

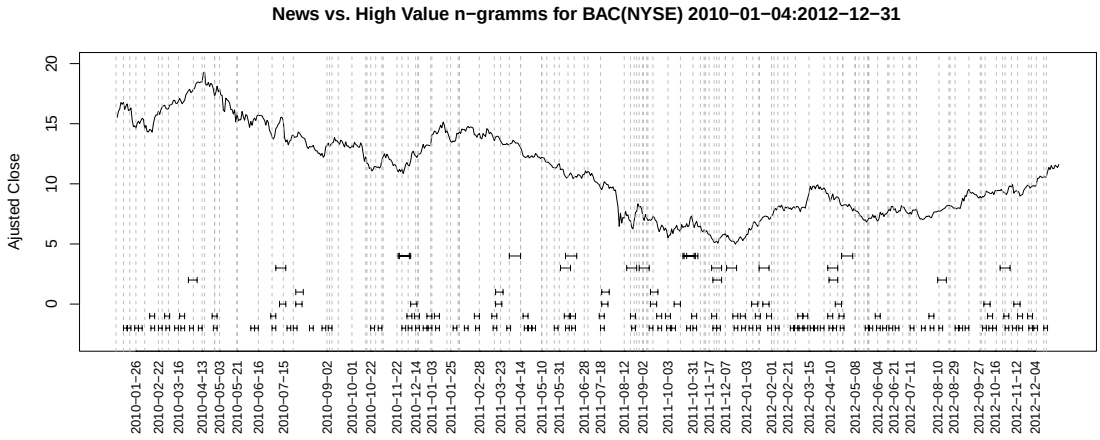

Fig. 4. Events and announcements of Bank of America Corp. together with the divergent n-grams. Horizontal bars represent non-unique divergent n-grams (5%-percentile). Vertical dotted lines are the official company events and announcements

The first observation is that the location of the longer divergent n-grams on the time axis does not necessarily coincide with the location of the shorter ones. In other words, no direct "filtration" of the divergent n-grams takes place. It may come from an instability of divergent n-grams against the noise of different origins. To mitigate this effect we may consider studying only stable divergent n-grams i.e. n-grams that possess filtration property. It is still unclear if such stable can be used as technical indicators marking anything of interest like markets top, beginning or end of a trend.

Another possible reason for the lack of filtration of divergent n-grams can be expressed as a hypothesis regarding the role of divergent n-grams in the indirect information exchange between market participants. The shorter ones are responsible for the shorter time range and are generated by interaction of short-horizon market participants. The longer ones are possibly related to the activity of longer horizon investors. The absence of the information exchange on the shorter

periods of time suggests that such exchange is also unlikely on longer time periods that contain the shorter one in question. Such ordinary periods of time, devoid of divergent n-grams, would represent "business as usual". Thus we may expect "filtration" of such ordinary stretches of time from smaller n to larger n.

2.5. Divergent n-grams and market events

A hypothesis that the divergent n-grams are connected to particular market events can be tested directly. The Fig. 4 shows an overlay of the official company events and announcements over price chart and divergent n-grams for ticker BAC. The events were pulled off the official investors section of the website of Bank of America Corp. The absence of direct connection between events and divergent n-grams appears to be inconclusive. The official events are polluted with those that are not really significant and may not include rumors, information leaks or external events significant for the company. The insider trading data published by the company may be also included into the analysis. It may be worth to attempt to filter out less significant official events and include important external ones in the further analysis.

Alternatively, one can judge the significance of an event by the information value of the n-gram that it generates (or is preceded by.) Both approaches require wider study on more diverse input data in order to rectify their methodology.

The divergent n-grams computed for one time window do not have to coincide with those computed for different but overlapping time window. This can be called time-shift (in-)stability. It turned out that divergent n-grams are usually reasonably stable. The Tab. 1 illustrates this observation with the data for the ticker BAC. The binary alphabet is $\{0, 1\}$ and represents up- and down- ticks.

Table 1. Non-unique divergent n-grams (5%-procentile) for two overlapping time windows. **I** stands for the period 2010-01÷2012-12; **II** stands for the period 2010-03÷2013-03; q is an n-gram

q	**I**	**II**	q	**I**	**II**
3	011		8	00110111	00110111
		111		11111000	11111000
4	0011	0011		10000111	10000111
				10001001	
5	11000	11000	9	111001000	111001000
6	110000	110000		101011100	101011100
7	1011111	1011111		010111001	010111001
	0110111			011100110	
		0100011			110010111

3. Discussion

The paper presents a novel approach of a study of the dynamics of financial series through the technique of frequency dictionary analysis. To do that, one has to quantize the original market data into a symbol series. Neither the quantization, nor a frequency dictionary approach themselves make a novelty; a combination of these two ideas makes that latter. It should be said that a continuous analog of the approach present above is also possible, due to wavelet technique. Meanwhile, the continuous case is more complex and brings less understanding of the issue standing behind the observed data. Nonetheless, the choice of an alphabet is of key point here.

The proposed approach to the analysis of time series based on the information capacity of n-grams extracted from the corresponding discrete texts can provide potentially valuable new tools and statistical metrics. The paper discussed several possible applications of the new approach and illustrated them with case studies of the actual market data. Also we discussed limitations of the approach and quantified them by estimating the noise limit and the aliasing effect. The connection of divergent n-grams with market events or their value as technical indicators remains a topic open for deeper investigation. However, even superficial observations with limited data and binary quantization hint at the connection of the divergent n-grams with implicit information exchange between market participants. Such hypothesis, if proven to be right, can have significant value for the market analysis.

References

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

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

[3] R.S.Tsay, Analysis of Financial Time Series. Financial Econometrics, Wiley & Sons, Inc, 2002.

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

[5] G.N.Pettengill, S.Sundaram, I.Mathur, The Conditional Relation between Beta and Returns, /it J. of Financial and Quantitative Anal., **30**(1995), no. 1, 101–116.

[6] M.D.Atchison, K.C.Butler, R.R.Simonds, Nonsynchronous securities trading and market index autocorrelation, *J. of Finance*, **42**(1987), no. 1, 111–118.

[7] Ya.Amihud, H.Mendelson, Trading mechanisms and stock returns: an empirical investigation, *J. of Finance*, **42**(1987), no. 3, 533–553.

[8] 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.

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

[10] 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.

Анализ финансовых временных рядов с помощью двоичных N-граммных частотных словарей

Михаил Г. Садовский,
Игорь Боровиков

Рассмотрена простейшая модель динамики временных рядов финансовых рынков для бинарной квантизации. Обсуждены наблюдаемые результаты и другие способы квантизации.

Ключевые слова: порядок, энтропия, условная энтропия, индикаторы.