Информационные процессы, Том 14, № 3, 2014, стр. 275–283.

SCOT stationary distribution evaluation for some examples

M. Malyutov*, P. Grosu*, and T. Zhang*

Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA E-mail: m.malioutov@neu.edu, pgrosu@gmail.com and zhang.tong@husky.neu.edu

Received September, 10, 2014

Abstract—We call Stochastic COntext Trees (abbreviated as SCOT) n-Markov Chains with every state of a string independent of the symbols in its more remote past than the **context** of **length** determined by the preceding symbols of this state. Previous somewhat confusing names for SCOT were VLMC, PST, CTW. We estimated SCOT parameters for testing homogeneity of data strings in No. 4, vol. 13 of this journal. Our more efficient SCOT fitting algorithm will be exposed elsewhere. Here, we *postulate* SCOT models and study their convergence without fitting SCOT from data sets. A SCOT *stationary distribution over contexts* is iteratively evaluated here explicitly in several examples. Our main tool is a 1-MC generated by the SCOT with the set of contexts as its state space.

KEYWORDS: Variable Memory Length Markov Chain, Stochastic Context Trees, Stationary Distribution of Contexts.

1. INTRODUCTION

Modeling random processes as full n-Markov Chains (n-MC) can be inadequate for small n, and over-parametrized for large n. For example, if the cardinality of the base state space is four, $n = 10$, then the number of parameters is around 3,15 millions. The popular Box–Jenkins ARIMA and Engel's GARCH in quality control and finance, are not adequate in applications to linguistics, genomics and proteomics, security, etc, where comparatively long *non-isotropic contexts* are relevant that would require huge memory size of the full n-MC.

The popularity of another tool – *sparse* Variable Memory Length MC (VLMC), has been increasing rapidly since a fitting algorithm was proved to be consistent for stationary mixing sequences in [5] and used for compression. In the 21st century, the predicted *inconsistent likelihood* of the VLMC model was first used in [1] for classifying genomics and proteomics highly nonstationary data.

Sparse SCOT over some alphabet A is a very special case of n-MC, where n is the maximal size of **contexts**. With a given string $x_{-n}, \dots, x_{-1}, x_0$ and small abuse of notation we define

$$C(x_0) = x_{-1}, \dots, x_{-k} := x_{-1}^{-k}, k \le n, \tag{1}$$

(to a current state x_0) is a subsequence of the preceding states x_{-1}^{-n} (see (1)) of the **minimal size** k such that the conditional probability

$$P(x_0|x_{-1}^{-m}) \equiv P(x_0|x_{-1}^{-k}), \forall m > k, k := |C(x_0)|. \tag{2}$$

is called the size of context $C(x_0)$. For large n, SCOT is sparse, if the total number of contexts $|\tau|$ is $O(n^k)$ for some fixed k. Note, that symbols of contexts are written in the natural order starting from the oldest one. The **same number** $|A|$ **of edges** goes from any node of the context tree other

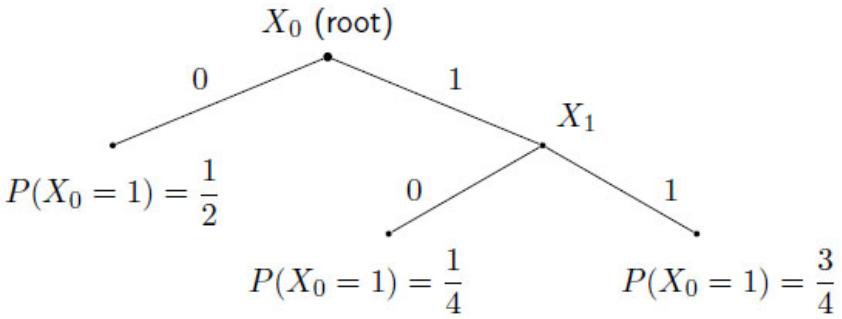

Fig. 1. The simplest stochastic context tree (Model 1).

than the context end (leaf). This fact called 'completeness' is crucial for our lemma in the next section.

The simplest SCOT model with alphabet $\{0, 1\}$, contexts $\{\mathbf{0}\}, \{\mathbf{01}\}, \{\mathbf{11}\}$ and corresponding transition probabilities $P(x_0 = 1)$ given preceding contexts are respectively $1/2, 1/4, 3/4$, as displayed above. Further, we find SCOT stationary distribution in this and several more models to prepare for more advanced topics: invariance, asymptotic normality and exponential tails of additive functionals which we expect to publish in the next issue of this journal. These are of interest after our statistical analysis in [4] showed an advantage of SCOT approximation over the popular GARCH in distinguishing between quiet and volatile regions of certain financial time series (FTS). The reason for this advantage seems to be an anisotropic memory required to predict FTS.

The SCOT stationary distribution evaluation implies derivation of its entropy rate according to (4.27) in [2] and the SCOT stationary distribution evaluation also lets us determine periodicity of SCOT and find transient contexts. Removing the latter can simplify SCOT based statistical inference and simulation as compared to [3]. SCOT models are usually consistently *estimated* from data and are thus only approximations of a SCOT describing the data explicitly. We finish with choosing parameters of our last SCOT model to match the historical Apple FTS.

2. 1-MC MODEL INDUCED BY SCOT

Before we expose our principal technical tool for SCOT, let us introduce it for the full n-MC. A full n-MC can be regarded as MC on the space of n-grams. Namely, by the chain rule:

$$P(x_0, x_1, \ldots, x_{n-1} | x_{-n}, x_{-n+1}, \ldots, x_{-1}) = \Pi_0^{n-1} P(x_i | x_{i-n}, \ldots, x_{i-1}). \tag{3}$$

Let T be a complete context tree and $C(T)$ be the set of its contexts.

To implement the same idea for SCOT models, it is sufficient to verify that after moving from x_i to $x_{i+1}, i = 0, \ldots,$ the context for the latter will **not include older than in** $C(x_i)$ **symbols**. This is formalized by the following

Definition. A context tree $C(T)$ is 'TailClosed' if for any string $z, |z| \geq \min_{t \in C(T)} |t|$, there exists $u \in C(T)$ and a substring $s, |s| \geq 0$, such that $z = \overline{su}$ meaning that z is a concatenation s, u in the natural ordering.

Let $M(T)$ denote the maximal difference between context sizes in $C(T)$.

Theorem. *If $M(T) \leq 1$, then the SCOT is TailClosed.*

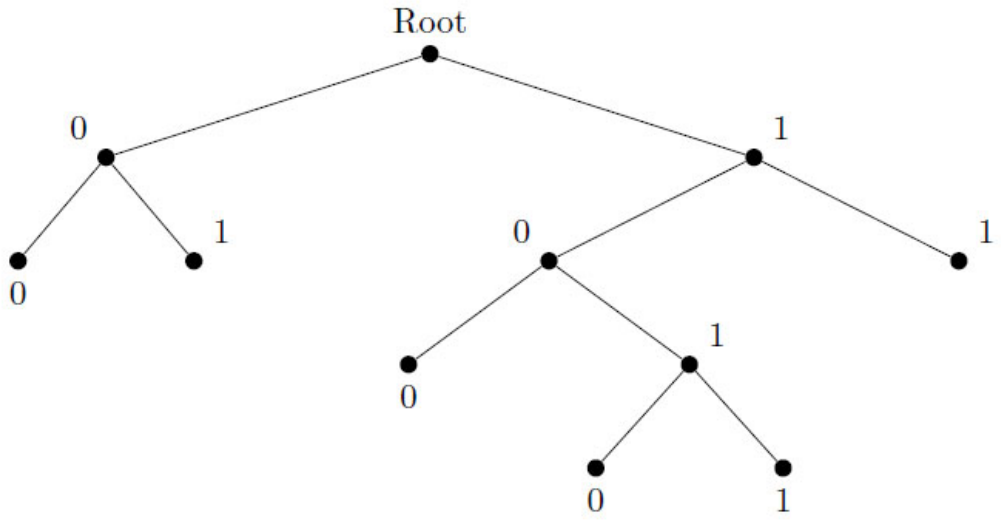

Fig. 2. Counterexample.

Note that every SCOT T must contain at least one complete set of 'siblings' with the largest context size, that is they are different only in the first symbol from the past. Denote by T' the tree obtained by replacing a set of 'siblings with one node t'.

The following auxiliary statements hold:

Lemma 1. $M(T') \leq 1 \leftrightarrows M(T) \leq 1$. Proof is straightforward.

Lemma 2. 1. *If $C(T)$ is TailClosed, then $C(T')$ is also TailClosed.*

Proof is exposed only for binary alphabet to simplify the notation. 1. Let $\overline{0X_1X_2X_m}$, $\overline{1X_1X_2\dots X_m}$ be siblings of the maximal size. Fix arbitrary string z such that its size $|z| \geq \min_{t\in C(T)} |t|$. There exists $t \in C(T)$, such that $z = \ at$. If $t \in \overline{0X_1X_2\dots X_m} \cup \overline{1X_1X_2\dots X_m}$, let $t' = \overline{X_1X_2\dots X_m}$. We have $z = \overline{a't'}$, and $t' \in C(T')$. If t is in the complement to $\overline{0X_1X_2X_m} \cup \overline{1X_1X_2\dots X_m}$, then $t \in C(T')$ itself.

Lemma 3. *Every binary context tree T such that $M(T) = 0$ is TailClosed.* Proof is straightforward.

Proof of the Theorem. Let the maximal context size be n. If $M(T) = 0$, our statement follows from lemma 3. If $M(T) = 1$, then there must be a sequence of binary context tree $T_1, T_2, \dots, T$ such that $M(T_1) = 0$ (say, we start with the full n-MC), and the maximal context size of T_1 is n; and T_{i+1} is obtained from T_i by cutting off the longest sibling leaves. By lemma 3, T_1 is TailClosed. By lemma 2, $T_1, T_2, \dots, T$ are all Tailclosed.

Counterexample. Consider $C(T) = \{00, 10, 001, 0101, 1101, 11\}$. If x_0=1 and $C(x_0)$=((10)), then (10)1=101, which is not a context.

Generally, the induced MC should be found by direct computations for particular SCOT.

3. STATIONARY DISTRIBUTION FOR MODEL 1

Thus we have the following

Definition. *The transition probability from $C(x_i)$ to $C(x_{i+1})$ equals to the transition probability from $C(x_i)$ to x_{i+1}.*

In our first example, the above definition leads to the following transition probability matrix **P** between contexts:

	11	**0**	**01**
11	0.75	0.25	0.0
0	0.00	0.50	0.5
01	0.25	0.75	0.0

This MC is ergodic and we evaluate its stationary distribution – row-vector $\mathbf{Q} = (q(\mathbf{11}), q(\mathbf{0}), q(\mathbf{01}))$ of contexts which is the unique solution to the equation $\mathbf{Q} = \mathbf{QP}$. This solution for this elementary example can be found by hand. In more advanced cases the following iterative procedure exists: use not bad preliminary guess $\mathbf{Q}^0$ and multiply $\mathbf{Q}^0$ by $\mathbf{P}^n$ for large n. Due to the well-known *exponential convergence rate* of these products to $\mathbf{Q}$ as $n \to \infty$, we approximate $\mathbf{Q}$ with arbitrary precision.

$\mathbf{Q}$ turns out to be $(1/4, 1/2, 1/4)$. Thus the entropy rate

$$H = -\sum_{i=1}^{3}\sum_{j=1}^{3} q_i P_{ij} \log_2 P_{ij} = 5/4 + 3/8 \log_2 (4/3).$$

4. MODELS 2

We start with over-simplified models to warm up and then move towards more realistic models. The idea behind these models is a caricature mimicking 'Galileo's inertia law' with friction (models 2), and 'Newton's second law of mechanics' with friction (model 3). All computer computations were made using the language **R** by the second author.

4.1. Ladder Model 2 (i)

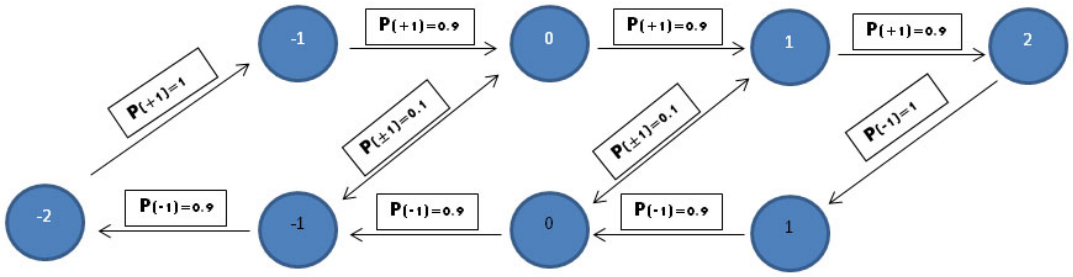

Fig. 3. Illustrated is ladder model 2(i).

The stationary distribution of model reveals an undesirable periodicity of the MC over contexts which we later fix by introducing a tiny positive probability of staying at every state.

$$X_n = \begin{cases} X_{n-1} + 1, \text{with prob } 0.9 \\ \qquad\qquad\qquad\qquad \text{if } X_{n-1} = X_{n-2} + 1 \text{ and } X_{n-1} \neq \pm l \\ X_{n-1} - 1, \text{with prob } 0.1 \end{cases}$$

$$X_n = \begin{cases} X_{n-1} + 1, \text{with prob } 0.1 \\ \qquad\qquad\qquad\qquad \text{if } X_{n-1} = X_{n-2} - 1 \text{ and } X_{n-1} \neq \pm l \\ X_{n-1} - 1, \text{with prob } 0.9 \end{cases}$$

$$X_n = \begin{cases} X_{n-1} + 1, \text{if } X_{n-1} = -l \\ X_{n-1} - 1, \text{if } X_{n-1} = l \end{cases}$$

Thus, the transition probability matrix is:

	-2.down	**-1.up**	**-1.down**	**0.up**	**0.down**	**1.up**	**1.down**	**2.up**
-2.down	0	1	0	0	0	0	0	0
-1.up	0.1	0	0	0.9	0	0	0	0
-1.down	0.9	0	0	0.1	0	0	0	0
0.up	0	0	0.1	0	0	0.9	0	0
0.down	0	0	0.9	0	0	0.1	0	0
1.up	0	0	0	0	0.1	0	0	0.9
1.down	0	0	0	0	0.9	0	0	0.1
2.up	0	0	0	0	0	0	1	0

The powers of transition probability matrix converge after even number of steps to:

	-2.down	**-1**	**0**	**1**	**2.up**
-2.down	0.25	0	0.5	0	0.25
-1	0	0.5	0	0.5	0
0	0.25	0	0.5	0	0.25
1	0	0.5	0	0.5	0
2.up	0.25	0	0.5	0	0.25

The powers of transition probability matrix converge after odd number of steps to:

	-2.down	**-1**	**0**	**1**	**2.up**
-2.down	0	0.5	0	0.5	0
-1	0.25	0	0.5	0	0.25
0	0	0.5	0	0.5	0
1	0.25	0	0.5	0	0.25
2.up	0	0.5	0	0.5	0

4.2. Model 2 (ii)

In Model 2 (ii), we modify the previous one by introducing a probability to stay at the same state to avoid periodicity:

$$X_n = \begin{cases} X_{n-1} + 1, \text{with prob } 0.8 \\ X_{n-1} - 1, \text{with prob } 0.1 \\ X_{n-1}, \text{with prob } 0.1 \end{cases} \quad \text{if } X_{n-1} = X_{n-2} + 1 \text{ and } X_{n-1} \neq \pm l$$

$$X_n = \begin{cases} X_{n-1} + 1, \text{with prob } 0.1 \\ X_{n-1} - 1, \text{with prob } 0.8 \\ X_{n-1}, \text{with prob } 0.1 \end{cases} \quad \text{if } X_{n-1} = X_{n-2} - 1 \text{ and } X_{n-1} \neq \pm l$$

$$X_n = \begin{cases} X_{n-1} + 1, \text{if } X_{n-1} = -l \\ X_{n-1} - 1, \text{if } X_{n-1} = l \end{cases}$$

The transition probability matrix is:

	-2.down	**-1.up**	**-1.down**	**0.up**	**0.down**	**1.up**	**1.down**	**2.up**
-2.down	0.1	0.9	0	0	0	0	0	0
-1.up	0.1	0.1	0	0.8	0	0	0	0
-1.down	0.8	0	0.1	0.1	0	0	0	0
0.up	0	0	0.1	0.1	0	0.8	0	0
0.down	0	0	0.8	0	0.1	0.1	0	0
1.up	0	0	0	0	0.1	0.1	0	0.8
1.down	0	0	0	0	0.8	0	0.1	0.1
2.up	0	0	0	0	0	0	0.9	0.1

The powers of the transition probability matrix converge to:

	-2.down	**-1**	**0**	**1**	**2.up**
-2.down	0.125	0.25	0.25	0.25	0.125
-1	0.125	0.25	0.25	0.25	0.125
0	0.125	0.25	0.25	0.25	0.125
1	0.125	0.25	0.25	0.25	0.125
2.up	0.125	0.25	0.25	0.25	0.125

The stationary distribution is $(0.125, 0.25, 0.25, 0.25, 0.125)$ which is $\left(\frac{1}{2n}, \frac{1}{n}, \ldots, \frac{1}{n}, \frac{1}{2n}\right)$ where $l = 2n$.

Model 2(ii)

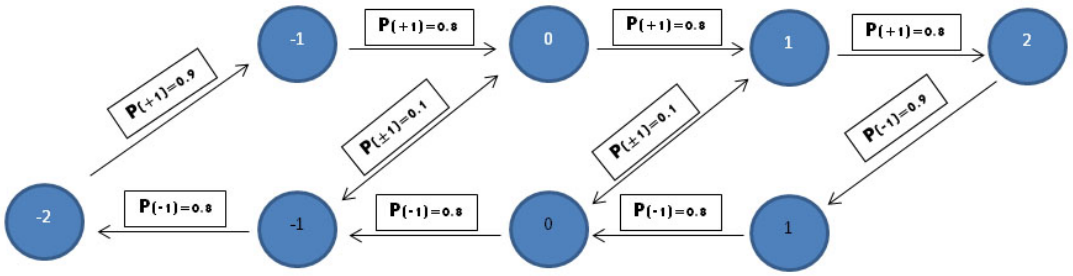

Fig. 4. Illustrated is model 2(ii).

5. MODEL 3

The contexts

5.1. Model 3 (i)

When $X_{n-1} \neq \pm l$

$$X_n = \begin{cases} X_{n-1} + 1, \text{with prob } 0.9 \\ \\ X_{n-1}, \text{with prob } 0.1 \end{cases} \quad \text{if } X_{n-1} = X_{n-2} + 1 \text{ or } X_{n-1} = X_{n-2} = X_{n-3} + 1$$

$$X_n = \begin{cases} X_{n-1} - 1, \text{with prob } 0.9 \\ \\ X_{n-1}, \text{with prob } 0.1 \end{cases} \quad \text{if } X_{n-1} = X_{n-2} - 1 \text{ or } X_{n-1} = X_{n-2} = X_{n-3} - 1$$

$$X_n = \begin{cases} X_{n-1} + 1, \text{with prob } 0.05 \\ X_{n-1} - 1, \text{with prob } 0.05 \\ X_{n-1}, \text{with prob } 0.9 \end{cases} \quad \text{if } X_{n-1} = X_{n-2} = X_{n-3}$$

When $X_{n-1} = \pm l, X_n = X_{n-1} - \frac{X_{n-1}}{abs(X_{n-1})}$

6. MODEL 3 DISTRIBUTION

Following the model described above and using an initial state $(x_1 = 0, x_2 = 1, x_3 = 2)$ of MC with $l = 15$ (totalling 31 states) we calculated X_{500} simulated 10,000 times. We got the following result:

-15	**-14**	**-13**	**-12**	**-11**	**-10**	**-9**	**-8**	**-7**	**-6**
0.0713	0.0349	0.0316	0.0243	0.0305	0.0288	0.0294	0.0258	0.0296	0.0248

-5	**-4**	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**
0.0265	0.0279	0.0297	0.0316	0.03	0.0315	0.0273	0.0292	0.0274	0.0253

5	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
0.0305	0.0296	0.0301	0.0302	0.0296	0.0294	0.0273	0.0287	0.0277	0.0389	0.0806

Below is a chart describing the above results:

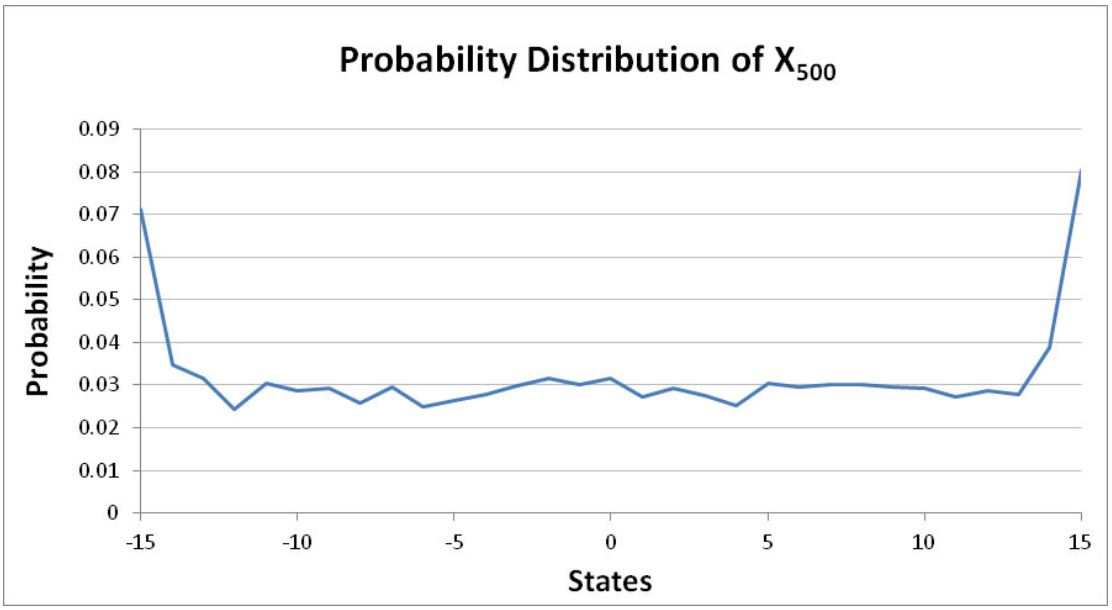

Fig. 5. Illustrated is the probability distribution of X_{500} for the different states.

It is plausible from this chart that the stationary distribution is uniform inside $[-l, l]$. To prove this, we introduce the row $2l + 2$-vector with all inner entries equal to 1 and two boundary entries x. Multiplying it by P from the right, we obtain the same vector, if x is chosen appropriately. This computation will be made by us in some time.

The stationary distribution really seems to be close to uniform sufficiently far from the ends of the interval judging by the above displayed simulation result.

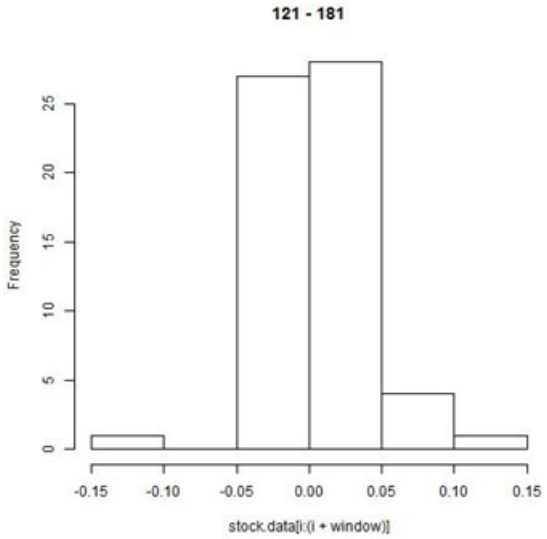

Fig. 6. Illustrated is 121–181.

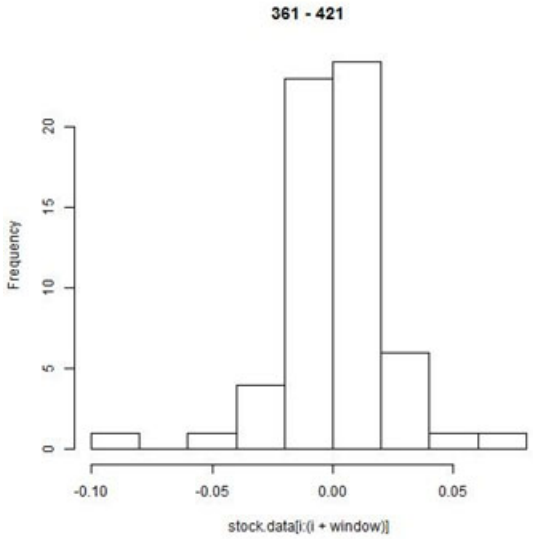

Fig. 7. Illustrated is 361–421.

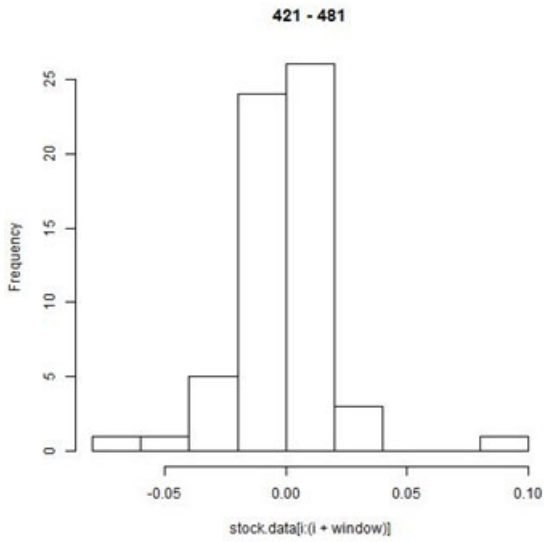

Fig. 8. Illustrated is 421–481.

7. SIMPLIFIED MODELING OF FINANCIAL DATA TO MATCH THE LADDER MODEL 3

Financial log-returns take arbitrary real values. Nevertheless, it might be worthwhile to model their 'infinitesimal evolution' as Ladder Model 3 with parameters matching the data. For doing this, we plot histograms of first and second differences of log-returns with appropriate bins. The Apple data was taken from a two-year period and then the difference of the difference of the log was performed on the adjusted closing price — this would be 2-nd derivative on the log of the closing price. Then the data was split into 60-day periods. Then the "centrality"of each 60-day window was checked via a histogram for frequency of variation. The day-window is the title of each histogram, which are shown above.

8. CONCLUSIONS, DISCUSSION AND FUTURE DIRECTIONS

Our induced 1-MC approach enables SCOT stationary distribution evaluations in interesting cases. The exposed over-simplified transparent models admit more practical generalizations. The 'Context' training SCOT software used in [4] has a severe restriction on the SCOT alphabet size. Thus we had to use drastic quantization of measurements. Our second author prepared the 'SCOT' parallel training software on multiple computers that remove this severe limitation of alphabet size and enables many additional options. We expect to generalize the Ladder Model 3 to many state levels instead of two, prove its convergence to a Brownian motion-like process under large l, small jump sizes and time intervals between SCOT jumps; develop stochastic equations with its respect and evaluate the European option prices according to this new approach.

REFERENCES

1. Bejerano G. *Automata learning and stochastic modeling for biosequence analysis*, PhD dissertation, Hebrew University, Jerusalem, 2003.

2. Cover T.M. and Thomas J.A. *Elements of information theory*, second edition, Hoboken: Wiley, 2006.

3. Garrivier A. Perfect Simulation Of Processes With Long Memory: A 'Coupling Into And From The Past' Algorithm, *Random Structures and Algorithms*, Arxiv Number: 1106.5971, October, 2014.

4. Malyutov M., Zhang T., Li Y., and Li X., Time series homogeneity tests via VLMC training.*Information Processes*, 2013, vol. 13, 4, 401–414.

5. Rissanen J. A universal data compression system, *IEEE Trans. Inform. Theory*, Vol. 29, No. 5, 1983, 656–664.