Информационные процессы, Том 17, № 1, 2017, стр. 61–69

SCOT Approximation and Asymptotic Inference I

M. Malyutov

Mathematics Dept., Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA E-mail: m.malioutov@neu.edu

Received March 16, 2017

Abstract—Approximation of stationary strongly mixing processes by SCOT models and the Le Cam-Hajek-Ibragimov-Khasminsky locally minimax theory of statistical inference for them is outlined. SCOT is an m-Markov model with sparse memory structure. In our previous IP papers we proved SCOT equivalence to 1-Markov Chain with state space — alphabet consisting of the SCOT contexts. For the fixed alphabet size and growing sample size, the Local Asymptotic Normality is proved and applied for establishing asymptotically optimal inference. We outline what obstacles arise for a large SCOT alphabet size and not necessarily vast sample size.

KEYWORDS: Strong mixing, Strongly stationary sequences, Local Asymptotic Normality, Local Asymptotic Minimaxity, SCOT models, Edgeworth expansion.

1. Introduction

Strongly stationary sequences as an object of advanced Probability theory are studied in the *first part* of this paper culminating in their LAN property, section 5.

Approximability of strong mixing sequences by m-MC with large m belongs to the mathematical folklore. In view of exponential complexity of general m-MC, ARMA-models were their popular surrogates until sparse memory m-MC was introduced in [21] for compression aims. We study conditions for sparse m-MC approximation of strong mixing stationary sequences in section 2.

The ergodicity of Markov Chains (MC) and Asymptotic Normality (AN) of additive state and transition functions of their paths has been subject of numerous studies starting from the pioneering works of A. A. Markov and S. N. Bernstein in the beginning of 20th century. Among many popular surveys—[4, 20, 24]. Statistical inference for MC has become popular after [1, 22]. The second of these references introduced the MC Local Asymptotic Normality (LAN) following the Le Cam-Hajek asymptotic locally minimax inference theory. An elementary exposition of this theory is in [24, 25]. The *traditional asymptotics* with *fixed MC* and growing sample size N is considered in all these references. Our section 5 outlines the simpler straightforward LAN derivation for finite MC with fixed alphabet following with some revisions [24, 25] rather than the longer way based on the CLT for MC reduction to the more general Martingale CLT which requires a rather cumbersome Poisson-like problem solution which is not straightforward ([20]).

Our statistical SCOT modeling ([17, 23]) of financial data discovered a small size of the context tree, while literary texts showed the number of SCOT contexts $m > 2000$; m is the size of 1-MC alphabet equivalent to SCOT, if the corresponding SCOT has a perfect memory [28], (in [18,23] another name:'TailClosed' is used for the same object). Otherwise, a perfect memory envelope of the original SCOT [28] should be used with even larger alphabet size.

The literary texts example in [17, 23] shows that the traditional asymptotics for deriving AN of additive SCOT functions can be inadequate. A similar change of asymptotics has been suggested by A. N. Kolmogorov for statistical classification of objects characterized by many normally distributed characters each, see [9], where Kolmogorov's ideas are exposed.

To illustrate what happens when both m and N grow simultaneously, we consider in subsection 3.1 the spectral decomposition of cyclic random walks nicely exposed in [10], pp. 377–378 and 434–435.

Our change of asymptotics uses the first order Edgeworth expansion for the additive functions (see [2, 3, 14, 16] among many publications). The principal multiplier $\mu_3\sigma^{-3}/\sqrt{N}$ of the residual term may grow with m which worsens the precision of approximation, see section 4. Here μ_3, σ are the stationary third moment and standard deviation of X_i respectively.

Sections 6-7 constitute the *mathematical statistics part* of this paper. We outline asymptotically optimal inference estimation and testing local hypotheses under LAN validity. In particular, the estimation of transition probabilities is a part of the change-point problem — distinguishing abrupt changes in SCOT model from its small deviation.

Section 7 justifies SCOT homogeneity testing results of [17, 23] in the framework of local (contiguous) alternatives theory under LAN. Testing very distant alternatives was exposed earlier in [17, 23] for an example of screening out active inputs of a multivariate regression model with colored noise using the large deviations probability results.

2. Approximation by SCOT

Approximability of strong mixing sequences by m-MC with large m belongs to the mathematical folklore and is widely used without rigorous definitions in the Information Theory, see [8]. We consider a strictly stationary discrete time process $X_t \in \mathcal{X}, -\infty < t < \infty$, $\mathcal{X}$ is the alphabet, with potentially infinite dependence

which can be approximated uniformly by an m-Markov chain (UA-m-MC condition).

By this we mean that for any $\varepsilon > 0$ there exists a positive integer $m(\varepsilon) > 0$ such that

$$|P(X_0|X_{-\infty}^{-1}) - P(X_0|X_{-m}^{-1})| < \varepsilon \; a.s.P$$

for any X_0 and past sequences $X_{-\infty}^{-1}$.

Remark 1. *Apparently, a uniform version of exponential memory decay absolute mixing (attributed to A.N. Kolmogorov in [26]) can guarantee a uniform approximation by an m-MC.*

Numerous conditions of strong mixing sequences are reviewed in [5]. Theorem 1.2 and Remark 3 in [6] assure that a very artificial stronger version of absolute regularity is equivalent to the fact that the sequence is m-MC for some m.

Assume now the UA-m-MC condition of a stationary sequence $x_{-\infty}^{\infty}$ and fix $m(\varepsilon)$ of the approximating m-MC which is assumed ergodic. Introduce m-gram—sequence x_1^m, and 2^m *contexts* for each of 2^m different realizations a_1^m of m-gram. The context to a_1^m is its final part of minimal length $l(a_1^m)$ such that the conditional distributions $P(X_{m+1}|x, a_{m-l}^m)$ do not depend on x up to joint error probability $< \varepsilon$. This statement is described by simultaneous validity of obvious $A \times (A-1)$ double inequalities. Not occurring m-grams are ignored. To streamline introduction, we assume that there are no such m-grams. Such a twice approximated stationary sequence will be called ε-SCOT with small abuse of notation.

Finally, the *memory spectrum* $\mathcal{M} = m_1^{2^m}$ is the 2^m-vector of context lengths along 2^m paths from the root to the past.

We combine all preceding developments into the following:

Definition. If a UA-m-MC has a memory spectrum $\mathcal{M}$, then $X_{-\infty}^{\infty}$ is an ε-SCOT with the corresponding context length distribution.

We say that $X_{-\infty}^{\infty}$ has a *sparse m-MC representation*, if the average prediction length satisfies:

$$2^{-m}\sum_{i=1}^{2^m} Em_i \ll m.$$

Remark 2. If under UA-m-MC condition, we are given a long string with a vast collection of m-grams, then the probability context length distributions can be replaced with their corresponding consistent frequencies. This allows the sparse m-MC training dealt with in [12, 15, 21]

Remark 3. Averaging context lengths over their stationary distribution gives a better sparsity indicator—the mean prediction length. The average prediction length is a preliminry indicator before the stationary distribution for the contexts is found. The median, or another quantile collection, or other functions of $\mathcal{M}$ over the stationary distribution can be also used for defining sparsity.

Remark 4. Another indicator of sparsity is the entropy rate of n-string which is much lower for the UA-m-MC with sparse memory than a general linear one. For example, the entropy rate of arbitrary n-string in the Comb model, [23], does not exceed a const.

Remark 5. Widespread sparse processes in nature phenomenon is explained by the 'Occam razor' or 'Bottleneck' popular philosophical principles.

3. Asymptotic Normality of additive transition functions

Our monograph [23], preceding IP publication [18] and especially [28] established the equivalence of a perfect memory sparse SCOT to 1-MC with state space consisting of the m-MC contexts which we call alphabet $\mathcal{A}$ of cardinality A . For not perfect memory sparse SCOT, its perfect memory sparse *envelope* studied in [28] plays this role. Thus,

by first applying UA-m-MC and $\mathcal{M}$ conditions, we reduced a stationary sequence to an m-MC with sparse memory structure, and now reduce it further to a 1-MC with alphabet $\mathcal{A}$.

We develop further asymptotic theory mostly for fixed A and large sample size and, therefore, for a finite ergodic MC avoiding arbitrary $\varepsilon > 0$ in previous approximations.

Let $X_i, i = 0, 1, \ldots,$ be the subsequent values of ergodic MC with alphabet $\mathcal{A} = \{1, \ldots, A\}$ and transition matrix $P = (p_{jk}, 1 \le j, k \le A); \pi = \pi(0)$ be the row-vector-initial distribution of X_0 and π^* be the stationary π. Finally, let $f(\cdot, \cdot)$ be a real function on $\mathcal{A} \times \mathcal{A}$. We call $S_n = \sum_{i=0}^{n} f(X_i, X_{i+1})$ an *additive transition function* (ATF) of MC X_i. An important ATF example is the *loglikelihood* $l_n(\theta)$ of a string X_0^n depending on vector θ of all transition probabilities and the *loglikelihood ratio* $r_n(\theta, \theta')$ which asymptotic representation in section 5 establishes the LAN property.

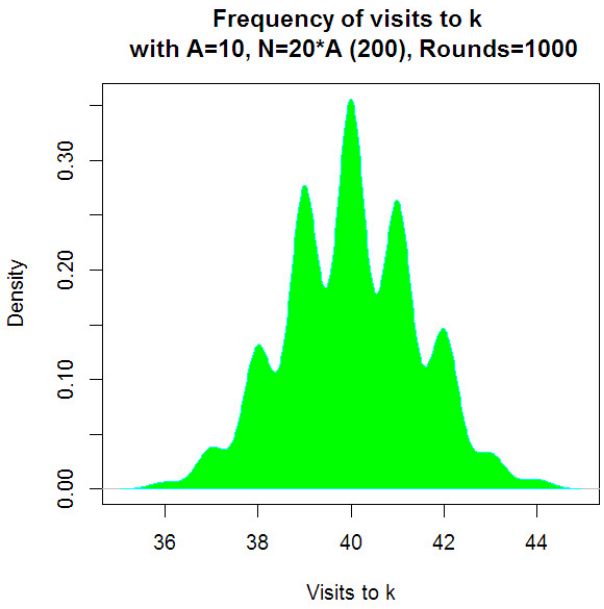
3.1. Ouverture: asymmetric cyclic RW example

To illustrate what happens when both A and N grow to infinity, let us consider the asymmetric cyclic random walk (RW).

The alphabet consists of equidistant circumference points $\exp(2ik\pi/A), k = 0, 1, \ldots, A-1$, i is the imaginary unit. The asymmetric cyclic random walk stays in the same state and jumps to the nearest left state with probabilities $1/2$. Introduce $\theta = \exp(2i\pi/A), s_r = ((1/2)(1 + \theta^r)$. Equation (2.11) of [10] establishes the power n of transition matrix spectral decomposition

$$p_{jk}^{(n)} = (A^{-1})\sum_{r=0}^{A-1} \theta^{r(j-k)} s_r^n. \tag{1}$$

We see from (1) that eigenvalues of the transition matrix are $O(A^{-1})$ apart as $A \to \infty$ which means that we cannot separate the maximal of them from the rest and restrict spectral expansion to just one 'maximal



Histogram

summand'. The term $p_{jk=0}^{(A/2)}$ corresponds to the additive state function for the indicator function of state $A/2$. Obviously, this function takes the value 0, if the initial state is 1 and number of summands less than $A/2$. Distribution of the sum is far from Normal, if few more summands are involved.

This fact is displayed in empirical histogram of visits to the state $A/2$ (fig.1), where the sample size N is 20 times more than A prepared by P. Grosu as a result of 1000 simulations. It shows several slightly intersected clusters far from the overall Normal histogram.

3.2. The AN of additive MC functions under fixed alphabet size

The most popular derivation of the CLT for MC nowadays is based on a reduction to the more general Martingale CLT which requires rather cumbersome approximations to the Poisson inverse-problem-like solution which is not straightforward (see e.g. [20]).

To compose parts of derivation together and draw attention of non-Russian-reading researchers especially, we outline a simpler approach based on results of the Russian Probability school.

Let **e** be A-column consisting of ones. For a real number t, introduce a new matrix $P(t)$ with entries $p_{jk}(t) = p_{jk}\exp(tf(j,k))$ and start with an elegant expression of S_n' moment generating function (MGF):

$$F_n(t) = E_\pi \exp(tS_n) = \pi P^n(t)\mathbf{e}. \tag{2}$$

The proof of insufficient for our aims particular case of $f(\cdot,\cdot)$ depending only on its second argument (additive state function (ASF)) is displayed in [24], pp. 230-232, and erroneously attributed there to A. A. Markov. The origin of this formula remains unclear to us. A. A. Markov actually used a cumbersome method of moments for deducing AN of S_n. We omit the detailed derivation of this formula. It is straightforward via sequential conditioning. At first $E(E(F_t|X_0^{n-1})) = F_{n-1}(\cdot)P(t)e$, then similar conditioning on X_0^{n-2}, etc.

To simplify further exposition, let us assume that all entries of $P = P(0)$ (and therefore also of $P(t)$) are positive. In view of ergodicity of $P = P(0)$, this is certainly valid for some power of $P = P(0)$, (see [10]), which is sufficient for our aims in this paper. Thus, $p_{jk}(t) = p_{jk}\exp(tf(j,k))$ is also strictly positive. The

Perron-Frobenius theorem implies that the isolated eigenvalue $\lambda(t)$ of $P(t), 0 \le t < \infty$, with the largest real part exists. Due to analicity of $P(t)$ and the theorem of implicit functions, this unique root $\lambda(t)$ of the equation

$$det(P(t) - \lambda I) = 0, \tag{3}$$

is an infinitely differentiable function of t in a neighborhood of $\lambda(0) = 1$. Attached to eigenvalue $\lambda(t)$ are row eigenvector $q_t \to \pi$ as $t \to 0$ and column eigenvector $e(t) \to e$ as $t \to 0$ infinitely smoothly depending on t, with unit scalar product. Then $P_1(t) = \lambda(t)e(t)q_t$ is such that $P(t) - P_1(t) = P_2(t)$ is exponentially smaller than $P_1(t)$ due to the Perron-Frobenius theorem. For our aims in this paper, $P_2(t)$ can be ignored.

Ergodicity: A normalized additive MC functions (ATF and ASF) shifted with time are obviously a stationary process converging to $\mu = E^* S_n/n$ as $n \to \infty$, where E^* refers to the stationary distribution π^*.

The proof (see [24], pp.236–237) of the AN of normalized additive ASF functions via applying twice the L'hospital's rule to its MGF is pretty standard given our representation of its MGF and similar to that in the IID case, see e.g. [13]. The proof in the ATF case is essentially the same. Of some interest is that the limiting distribution under standard normalization can be singular due to the null limiting variance.

A simple example of such anomaly for additive state function is the *symmetric cyclic RW* with four states and equally likely transitions to each of two neighbors, and alternating ± 1 function between neighboring states. Values ± 1 necessarily alternate also in time killing each other. Thus $S_{2n} = 0$, while $S_{2n+1} = \pm$ for all n and the standard $1/\sqrt{n}$ normalization provides the limiting null variance.

4. The Edgeworth expansion of additive MC functions under fixed alphabet size

Asymptotic expansion of additive functions appeared in [2,3], [16], [14] under various conditions which certainly include the case of a duly smoothed finite ergodic MC.

A simple appropriate smoothing procedure is described in theorem 2, XVI,4, [11].
In [16] the *first terms* of asymptotic expansion under Cramer-type conditions are:

$$P(N^{-1/2}(\sum_{i=1}^{N} f(x_i) \le x)) = \Phi_\sigma(x) + \phi_\sigma(x)(\mu_3/\sigma^3)q(x)N^{-1/2} + O(N^{-1}).$$

Here ϕ and Φ are PDF and CDF of the central Normal RV with StD σ, q is expressed in terms of the first Hermite polynomial $1 - x^2$.

The principal multiplier $(\mu_3/\sigma^3)/\sqrt{N}$ of the residual term may grow with m which worsens the precision of approximation. Here μ_3, σ are the stationary central third moment and standard deviation of X_i respectively.

In particular, for our circular MC of section 3.1 and the indicator function of state $A/2$ as an example of additive function, the mean is $A^{-1}, \mu_3 = A^{-1} + O(A^{-2}), \sigma^2 = A^{-1} + O(A^{-2})$. Thus, the principal multiplier of the residual $O(A^{1/2}/\sqrt{N}) \to 0$ only if $(A/N) \to 0$ as $N \to \infty$.

5. The Local Asymptotic Normality of SCOT under a fixed context cardinality

One of our principal aims is to outline the *Local Asymptotic Minimaxity* (LAM) and the Locally Asymptotically Most Power (LAMP) of the likelihood based inference and of its certain approximations. The LAM in parameter estimation as formally defined further in section 6, means that the deviation of the estimate from the true parameter value θ^* is as minimal as possible in the local minimax sense.

It is implied by the LAN condition (see e.g. [22, 25]) for MC which will be introduced immediately. The principal role in the LAN proof is played by the AN of the ATF functions (established in section 4 and missing in [25]). Two elementary corollaries of ergodicity ending the LAN proof and best exposed in [25] are omitted here.

The Local Asymptotic Normality (or simply LAN) introduced in Le Cam (1960) is the following decomposition of

$$r_n(\mathbf{u}) = \ln[P_{\theta+n^{-1/2}\mathbf{u}}((X_0^n))/P_\theta((X_0^n))], \mathbf{u} \in \mathbf{R}^A :$$

$$r_n(\mathbf{u}) = u^T\lambda - (1/2)\mathbf{u}^T J\mathbf{u} + \psi_n(\mathbf{u}),$$

where

$$\lambda \sim N(0, J), J = E_{\pi^*}\partial r(\theta)\partial r(\theta)^T$$

is the limit of the mean in the ATF $r(\cdot)$ Jacobian multivariate AN approximation and $\psi_n(\mathbf{u})$ converges in $P_\theta(X_0^n)$-probability to zero.

This expansion for a univariate parameter via the Taylor expansion of the second order is proved in [25] referring to the much more involved exposition in [22] for the AN proof of the ATF in general case under standard regularity conditions.

The uniformity of the residual $\psi_n(\mathbf{u})$ convergence in $P_\theta^{(n)}$- probability to zero can be proved by the more elegant Lagrange-type integral representation of the second order residual in the Taylor expansion as in [19]. Namely, for all $K > 0, a > 0$

$$\lim_{n\to\infty} P_{\theta+n^{-1/2}\mathbf{u},\sup||\mathbf{u}||<K}(|\psi_n(\mathbf{u})| > a]) = 0.$$

6. The Local Asymptotic Minimaxity of Likelihood-Ratio-like tests under a fixed alphabet size

Let the distribution family P_θ satisfy LAN condition in $\theta = \theta^*$ with the identity Fisher information matrix, $||\cdot||$ be the Euclidean norm. A function $w(\cdot) : \mathbf{R}^p \to \mathbf{R}^+$ is called bowl-shaped if $\{\mathbf{u}|\ w(\mathbf{u}) \le a\}$ are closed bounded symmetric convex sets for any $a \ge 0$. An increasing continuous bowl-shaped function $w(\cdot) : \mathbf{R}^+ \to \mathbf{R}^+, w(0) = 0$, is called a loss function.

The fundamental Hajek's lower bound for the LAM-risk of any estimate T_n for any loss function $w(\cdot)$ and $\delta > 0$:

$$\liminf_{n\to\infty} \sup_{||\theta-\theta^*||<\delta} E_\theta w(n^{1/2}||T^n - \theta||) \ge \int (2\pi)^{-1/2} w(\mathbf{u}) e^{-|\mathbf{u}|^2/2} d\mathbf{u},$$

holds. In general, the positively definite Fisher information J determines the norm in the risk function definition.

The LAM property of the Maximum Likelihood (ML) estimate and of the Fisher score update to ML given a qualified consistent prior estimate for θ are exposed in [22, 25]. [19] shows sufficiency of a usual consistent estimate for θ for LAM of the Fisher score update given the uniform LAN property.

The third Le Cam's lemma ([7, 22]) proves that the AN of a statistic under the null hypothesis implies its AN under the alternative distribution under contiguous distribution families and the LAN condition.

7. Locally asymptotically optimal tests

The most transparent overview of the Locally Asymptotically Most Powerful (LAMP) tests under LAN condition for I.I.D samples is in [7]. Given LAN property, it differs insignificantly from the one for MC in [22].

The main distinction of the LAMP approach originated in Le Cam's works from the traditional one, is that the 'close' alternatives $\mathbf{u}(n^{-1/2})$ are considered for the sample size $n \to \infty$. This enables limiting positive significance level and power asymptotically and a transparent application of the familiar testing shift theory for multivariate Normal. We now give schematic simplified overview of this theory following [7] and shortening our notation for transparency in an obvious way.

The Neyman-Pearson lemma gives the most powerful test of significance level α against alternative $\mathbf{u}(n^{-1/2})$ as

$$r_n(\mathbf{u}) = \ln[P_{\theta+n^{-1/2}\mathbf{u}}((X_0^n))/P_\theta((X_0^n))] > C_{n,\alpha},$$

with parameter $C_{n,\alpha}$ determined from equation $P_{n,0}(r_n > C_{n,\alpha}) = \alpha$.

The LAN condition converts this into the asymptotic equality $C_{n,\alpha} = z_\alpha\sqrt{J}\mathbf{u} - \mathbf{u}^T J\mathbf{u}/2$ which is equivalent to

$P_{n,0}(r_n < x) \to \Phi((x + \mathbf{u}^T J\mathbf{u}/2)/\sqrt{\mathbf{u}^T J\mathbf{u}})$

The power $\beta_{n,\mathbf{u}} = P_{n,\mathbf{u}}(r_{n,\mathbf{u}} > C_{n,\alpha})$ as $n \to \infty$ implying

$P_{n,\mathbf{u}}(r_n < x) \to \Phi((x - \mathbf{u}^T J\mathbf{u}/2)/\sqrt{\mathbf{u}^T J\mathbf{u}})$

Thus, $\beta_{n,\mathbf{u}} = P_{n,\mathbf{u}}(r_n > z_\alpha\sqrt{J}\mathbf{u}) \to 1 - \Phi(z_\alpha - \sqrt{J}\mathbf{u}) = \Phi(\sqrt{J}\mathbf{u} - z_\alpha)$ which means (see e.g. [7], (8.1.19)) that the limiting asymptotic power of our test is asymptotically maximal for every given alternative $\mathbf{u}$ in view of the Neyman-Pearson lemma. Thus, our test is LAMP.

Let us apply the preceding theory to the homogeneity of multivariate distributions of the large strongly stationary ergodic training string T and a query string Q. We use the nonparametric test of [17].

The first stage is estimation of the SCOT model of the string T following the algorithm in [15]. We refer to this publications for the details.

We assume

1. The T's and Q's good approximability by a sparse SCOT and

2. fulfillment of the LAN condition for the equivalent 1-MC over their contexts.

We cut the query string into K slices of the same length. Then, using the SCOT model of T we find the loglikelihoods $L_Q(k)$ of query slices Q_k and of strings S_k simulated from the training distribution of the same size as $Q_k, k = 1, \ldots, K$, (for constructing simulated strings, see e.g. algorithm in [15]).

We then find log-likelihoods $L_Q(k)$ of Q_k, $L_S(k)$ of S_k using the derived probability model of the training string and the average $\bar{D}$ of their difference D which approximates the likelihood ratio statistic discussed above. The averaging over slices is used for empirical evaluation of the loglikelihood variances since our testing homogeneity problem is completely nonparametric.

We assume though that the multivariate distributions of the training and the query strings are contiguous. In particular, for literary applications this assumption means that both texts are written in the same language, and admissibility of texts is the same for T and Q.

Next, due to the asymptotic normality of log-likelihood increments both for the null hypothesis and alternative (third LeCam's lemma), we can compute the usual empirical variance V of $\bar{D}$ and the t-statistic t as the ratio $\bar{D}/\sqrt{V}$ with $K - 1$ degrees of freedom (DF). We find K^* from the empirical condition that $t(K^*)$ is maximal. Then, the p-value of homogeneity is evaluated for the t-distribution with $K^* - 1$ DF.

8. Discussion, open problems and acknowledgments

Our presentation on modeling and asymptotic inference of strongly mixing stationary sequences differs drastically from the material presented in traditional courses on stationary processes and connects this discipline with the classical MC-theory. Our AN derivation for ATF seems much more transparent than the traditional one.

A formalization of convergence of strongly mixing stationary sequences to m-MC remains the main task to clarify.

Another one is the relation of the memory-spectrum and the entropy-based approaches for characterizing the sparsity of approximating m-MC.

The main challenging problem is to prove accurate asymptotic results for the case of rising alphabet of MC. simultaneously with the sample size.

The author is grateful to P. Grosu for programming and running simulation and to V. N. Tutubalin for the useful discussion.

References

1. Billingsley P. *Statistical inference for Markov chains*, University of Chicago Press, 1961.
2. E. Bolthausen E. *The Berry-Esseen theorem for functionals of discrete Markov chains*, Z. Wahrsch. Verw. Gebiete, **54**, 1980, 59-73.
3. Bolthausen E. *The Berry-Esseen theorem for strongly mixing Harris recurrent Markov chains*, Z. Wahrsch. Verw. Gebiete. **60**, 1982, 283-289.
4. Borovkov A. A. *Ergodicity and stability of stochastic processes*, Wiley, 1998.
5. Bradley R.C. Basic properties of strong mixing conditions. A survey and some open questions, *Probability Surveys*, **2**, 2005, 107–144.
6. Bradley R.C. A caution on mixing conditions for random fields. *Statist. Probab. Letters* **8**, 1989, 498–491.
7. Chibisov D. M. Lectures on the asymptotic theory of rank tests, *Lecture Notes NOTs 14. M.: Matematicheski Institut im. V. A. Steklova, RAN*, 2009, In Russian.
8. Cover T. M. and Thomas J. A. *Elements of information theory, second edition*, Hoboken: Wiley, 2006.
9. Deyev A. D. Asymptotic expansion of classification statistics in normal case, *Proceedings of the USSR Academy*, **195-4**, 759–776.
10. Feller W. *An introduction to Probability theory and its applications, volume 1, third edition*, Wiley, N. Y., 1967
11. Feller W. *An introduction to Probability theory and its applications, volume 2, second edition*, Wiley, N. Y., 1970.
12. Galves A. and Loecherbach E.. Stochastic chains with memory of variable length, In: *Festschrift in Honor of Jorma Rissanen on the Occasion of his 75th Birthday*, Tampere, TICSP series No. 38, Tampere Tech. Uni., 117–134, 2008.
13. Grinstead and Snell, *Introduction to Probability*, AMS, 2006.
14. Jensen J. L. Asymptotic Expansions for Strongly Mixing Harris Recurrent Markov Chains, *Scandinavian Journal of Statistics*, **16-1**, 1989, 47–63.
15. Mächler M. and Bühlmann P. Variable Length Markov Chains: methodology, computing, and software, *Journal of Computational and Graphical Statistics*, **13-2**, 2004, 435–455.
16. Malinovsky V. K. On limit theorems for Harris Markov chains. I. *Theory Probab. Appl., (English translation)*, **31**, 1987, 269–285.

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

18. Malyutov M. B. and Zhang T. Limit theorems for additive functions of SCOT trajectories, *Information Processes*, **15-1**, 2015, 89–96.

19. Malyutov M. and Protassov R. LAN and LAM, Convergence of Iterative Estimates and Optimal Design in Gaussian One-Way Mixed Model, *Journal of Statistical Planning and Inference*, **100-2**, 2002, 249–279.

20. S. P. Meyn and R. L. Tweedy *Markov chains and stochastic stability*. Springer, 1993.

21. Rissanen J. A universal data compression system. *IEEE Trans. Inform. Theory*, **29-5**, 1983, 656–664.

22. Roussas G. *Contiguity of probability measures: some applications in statistics*, Cambridge University Press, 1972.

23. Ryabko B., Astola J. and Malyutov M. *Compression-Based Methods of Statistical Analysis and Prediction of Time Series*, Springer International Publishing AG Switzerland, 2016.

24. Tutubalin V. N. *Probability and random processes theory. Mathematical foundations and applications*, Moscow State University Press, 1992 (In Russian).

25. Veretennikov A. *Parametric and nonparametric estimation for Markov Chains*, Moscow State University Press, 2000 (In Russian).

26. Volkonskii V.A. and Rozanov Yu.A. Some limit theorems for random functions I. *Theor. Probab. Appl.* **4**, 1959, 178-197.

27. Wefelmeyer W. Efficient estimation in Markov Chain models: an introduction, *Asymptotics, Nonparametrics and Time Series (S. Ghosh, ed.)*, 427–459, Statistics Textbooks and Monographs No. 158, Dekker, New York, 1999.

28. Zhang T. Perfect Memory Context Trees in time series modeling, *arXiv:* 1610.08910v1 [cs.LO], 2016.