Информационные процессы, Том 19, № 1, 2019, стр. 60–77

Quadratic risk of Change Point detection

M. Malyutov

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

Received 05.06.2018

Abstract—This paper continues my paper in IP, **17:3** 2017. We establish sharp order of magnitude bounds in Quadratic risk of Change Point detection online and offline for slow HMM-SCOT model.

KEYWORDS: Change Point detection online and offline, quadratic risk, martingale, submartingale, supermartingal, functional CLT, large deviations.

1. INTRODUCTION

Stochastic COntext Tree (abbreviated as SCOT) is m-Markov Chain (m-MC) 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. A parallel super-fast fitting and asymptotically optimal inference in a **sparse** SCOT model including the nonparametric homogeneity test are described in our previous papers. An alternative apparently less powerful and flexible model is Markov Chain of Conditional Order, see [17].

[28] and finally [35] established the equivalence of a **perfect memory sparse** SCOT to 1-MC with state space consisting of the collection of m-MC contexts which we consider as the new **alphabet** $\mathcal{A}$ of cardinality A. For not perfect memory sparse SCOT, its perfect memory sparse *envelope* (also studied in [35]) plays this role. The perfect memory condition *does not depend on prediction probabilities* assigned to the leaves of SCOT.

Thanks to above SCOT perfect memory reduction to a MC with enlarged state space, a substantial part of the present paper deals with statistical properties of 1-MC. Evaluation of log-likelihoods under SCOT requires **sophisticated software** and **cumbersome calculations**. The **statistical theory** of this processing is *miraculously* supported by the **classical 1-MC theory** with somewhat larger alphabet size A under perfect memory.

Markov *regime switching* models remain enormously popular in speech recognition, economics, finance, etc. Nonparametric segmentation in switching models without probability assignment of jump moments is in many papers by Brodsky and Darkhovsky. We model all regimes as long SCOT strings. Our segmentation method is a combination of preliminary online change point detection with its subsequent offline Maximal Likelihood update.

1.1. HMM model for speech recognition

Speech is modeled (Baum et al) as a sequence of **emissions**—phonems— random variables x_i. Elements of observed sequence x_i **depend only on current hidden letters of text** z_i which is modeled as a Markov Chain.

We refer to z_i as hidden states, see the Figure below. **Inference** about the parameters of the model and hidden MC uses only the observed $x_i, i = 1, \dots$.

The fast Baum-Welsh fitting HMM parameters has been successfully applied to speech recognition [25]. Its application to Genome modeling [11,34] by assigning IID emissions to the same part of Genome seems not justified. Markov switching models generalize HMM by considering **parametric** regimes ([9,16]). Nonparametric segmentation in switching models without probability assignment of jump moments is in many works by Brodsky and Darkhovsky, see [6]. We develop a model of **slow HMM with SCOT emissions (SCOT-HMM)** which seems a more realistic model for Genome, economics, analysis of combined authorship of literary works, or financial time series with piecewise volatility. We discuss in [20,23] well-SCOT-approximation of **mixing sequences**. Thus, mixing emissions are actually considered.

Our segmentation method is a superposition of preliminary **online change point detection** with its subsequent offline Maximal Likelihood update. The online part is a nontrivial modification of IID case in [18] using alternative risk function to that in [29].

1.2. Slow HMM-SCOT emissions model

We call HMM z_i **SLOW**, if mean time that HMM keeps staying in the same state is proportional to a large parameter l in all states, while the **sample size** is $kl, k \to \infty$. Emissions shown in dark in the following Fig. are modeled as **STRINGS of MC over the space of contexts** with transition matrix depending on the current HMM state.

Remark. Our HMM-SCOT model has nothing in common with VLMCHMM which analyzes independent emissions under a SCOT model for HMM [10].

The emissions MC x_i over the alphabet of SCOT contexts are assumed **ergodic, different for all states of HMM**, expectations are taken everywhere under their stationary distributions.

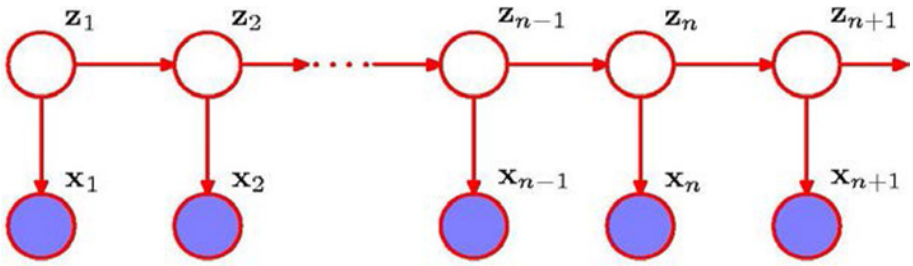

Fig. 1. HMM-SCOT scheme.

2. AUXILLIARY MATERIAL

Large Numbers Law (LLN) for additive **state** functions were first derived in [22], 1906. The MGF derivation of exponential convergence rate in LLN for more general additive **transition functions** (ATF) and their asymptotic Normality (AN) use the powerful classical

Theorem of Perron (PT). Any quadratic matrix with all entries positive has a positive simple eigenvalue R called its spectral radius, which is strictly greater than the moduli of the rest of the spectrum, R's corresponding eigenvector has all positive coordinates.

2.1. MGF, exponential rate in LLN and Asymptotic Normality of ATF

Let $\mathbf{1}_B$ be B-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):

$$F_n(t) = E_\pi \exp(tS_n) = \pi P^n(t)\mathbf{1}. \tag{1}$$

The proof with small gaps of insufficient for our aims particular case of $f(\cdot,\cdot)$ depending only on its second argument (additive state function (ASF)) is displayed in [30], 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 $E(E(F_t|X_0^{n-1})) = F_{n-1}(\cdot)P(t)\mathbf{1}$, then similar conditioning on X_0^{n-2}, etc.

Remark. *Another proof of (3) generalizes [32] for ATF. Introduce operator family mapping real functions $h(\cdot)$ on $\mathcal{B}$ into similar ones:*

$$T(t)h(x) = E_x(\exp[tf(x, X_1)]h(X_1)). \tag{2}$$

We have

$$MGF_x^{(2)}=E_{X_1,X_2}\exp[t(f(X_1)+f(X_2))]h(X_2)=E_{X_1}\exp[tf(X,X_1)](E_{X_2})h(X_2)\exp[tf(X_1,X_2)]=$$

$$= E_{X_1}\exp[tf(X,X_1)]T(t)_{X_1} = T^2(t)h(\cdot).$$

Putting $h(\cdot)\equiv 1$, we get (3) for $n=2$ and arbitrary initial x. Extension to arbitrary n is by induction.

The Perron theorem implies that $MGF_x \to MGF_\pi$ as $n\to\infty$.

$MGF_x^{(n)}$ is a *convex function of t* as a linear combination of exponentials with nonnegative coefficients.

To simplify further exposition, we 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 [12]), which is sufficient for proving our asymptotic results. Thus, $p_{jk}(t) = p_{jk}\exp(tf(j,k))$ is also strictly positive. The PT implies **isolated maximal eigenvalue**-spectral radius $R(t)$ of $P(t), 0 \le t < \infty$ existence. Due to analicity of $P(t)$ and the theorem of implicit functions, this unique root $R(t)$ of the equation

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

is an analytic function of t in a neighborhood of $R(0) = 1$. Attached to eigenvalue $R(t)$ are row eigenvector q_t and column eigenvector $e(t) \to \mathbf{1}$ as $t \to 0$ infinitely smoothly depending on t, with **unit scalar product**. This follows from the fact that each of $e(t) \to \mathbf{1}$ and q_t are the solutions of non-degenerate linear system of equations with the same non-degenerate minor of $P(t)$. Then

$P_1(t) = R(t)e(t)q_t$ is such that $R(t)^{-n}[P^n(t) - P_1^n(t)] := R(t)^{-n}P_2^n(t)$ is **strictly smaller** in matrix norm than $\exp(-an)$ for some $a > 0$ due to PT. Similarly, $F_n(\cdot) = \sum F_n^{(i)}(\cdot)$.

Remark. A standard Linear Algebra result implies existence of an invertible transformation $Q = Q(t)$ such that $Q^{-1}P(t)Q = \Lambda(t)$ admits the Jordan form decomposition with diagonal element $R(t)$ and an additional term $\tilde{\Lambda}(t)$. Thus $Q^{-1}P^nQ = Q^{-1}PQQ^{-1}PQ\times\cdots Q^{-1}PQ$ admits the Jordan form representation with diagonal $R^n(t)$ and additional term $\tilde{\Lambda}^n(t)$. As a result, $P^n(t) = P_1^n(t) + P_2^n(t)$.

$F_n^{(2)}(t)$ is shown to be negligible in our asymptotics derivation of $F(t)_n, n\to\infty$.

Introduce $H_n(t) = n^{-1}\log E_\pi \exp tS_n = [n^{-1}\log F^1(t)] + o(1)$.

Due to analycity of $r(t)$, $H_n'(0)/n = \mu_n \to \mu$ which is the limiting mean of ATF. Indeed, it equals

$$R'(0) + d/dt([q(t)\mathbf{1}][\pi e(t)])|_{t=0} + d/dt[\pi P_2^n(t)\mathbf{1}]|_{t=0}. \tag{4}$$

The second summand is obviously bounded, while the third one is bounded from above by

const $n\sum_{k=1}^{n-1}\exp(-ka)\exp(n-k)a \le const\exp(-a[n])$.

2.2. Asymptotic Normality

A normalized ATF shifted with time is obviously a stationary process converging to $\mu = \lim ES_n/n$ as $n \to \infty$. [30], p. 234, shows for ASF that $\mu = E_\pi f(X_1) = F_1'(t)$ at $t = 0$.

A similar result holds for a general ATF.

To prove the weak convergence to the limiting Normal approximation (possibly singular) under usual $\sqrt{n}$ normalization for **centered** ATF, we evaluate the first and second derivative of its normalized 'reduced' MGF $F_1^0(t)$ at $t = 0$.

The latter boils down to

$$d^2/dt^2[R^n(t)q(t)\mathbf{1}[\pi e(t)])]|_{t=0} + o(1). \tag{5}$$

Terms involving the first derivative $R'(0)$ of the centered normalized $P_1^0(t)$ vanish at $t = 0$ due to centering, say, $F'(0) = 0, F''(0) = \sigma^2$, we assume that $\sigma > 0$, see details in [30]. Only one term remains after neglecting as in the preceding proof exponentially small terms involving $P_2(t)$:

$$[(\pi(t/\sqrt{n})e)(\pi^* e(t/\sqrt{n})) + o(1)][1 + (t\sigma)^2/2n]^n \to \exp((t\sigma)^2/2)$$

as $n \to \infty$. This finishes the proof according to the classical Probability approximation theorems since the limiting MGF is that of the centered Normal distribution with variance σ^2.

Remark 1. We use further a **multivariate version** of the above AN theorem which proof generalizes naturally the above univariate case. The principal multivariate example is the log-likelihood ratio vs. the vector of alternatives. The covariance matrix J of the limiting Normal distribution replaces σ^2 in the above statements.

Remark 2. 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. [24, 32]). The proof (see [30], 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 rather similar to that in the IID case, see e.g. [15]. Our proof for ATF based on (3) is essentially the same.

Of some interest is that the limiting distribution under standard normalization can be singular due to the null limiting variance.

As a consequence, in this case there is no need for $\sqrt{n}$ normalization, and the residual distribution is bounded.

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 $\pm$function between neighboring states. Values ± 1 necessarily alternate also in time killing each other. Thus $S_{2n} = 0$, while $S_{2n+1} = \pm 1$ for all n and the standard $1/\sqrt{n}$ normalization provides the limiting null variance.

2.3. Martingale lemmas for log-likelihood

Likelihoods $\Pi_n = \Pi_{i=1}^n P(X_i|X_{i-1}), i = 1$, are martingales with respect to σ-algebras $\mathcal{F}_n$ generated by MC observations $X_i, i = 0, \ldots, n$, see [29]. Log-likelihoods $L_n = \sum_{i=1}^n l_{i,i-1}, l(X_i, X_{i-1}) = \log P(X_i|X_{i-1})$ are super-martingales as concave functions of martingales (see [29]) and thus admit the Doob-Meyer decomposition $L_n = S_n + r_n$ with martingale $m_n = l_n - \sum_{i=1}^n \sum_k P_{X_{i-1},k} l(k, X_{i-1})$, while *compensators* $r_n = sum_{i=1}^n \sum_{k=1}^A k P_{X_{i-1},k} l(k, X_{i-1})$ are $\mathcal{F}_{n-1}$-measurable ('predictable').

Further, $\exp tS_n$ as a convex function of S_n is a submartingale for $t \in \mathbf{R}$.

2.4. Exponential rate of Ergodicity

Here we derive functional exponential bounds for both martingale and compensator parts of the log-likelihoods for application in Change Point detection. 1. The functional version of the ergodic theorem for the compensator $r(X_k)$ is the following:

If we consider its latest k of $j = n + k$ summands Q_j, then due to a long past of length n the underlying distribution of X_1 can be taken as π.

Lemma 5.4. If $k = O(\log n)$, then absolute deviations of compensator Q_j from its mean jH exceed an arbitrarily chosen $\varepsilon > 0$ only a finite number of times.

Proof. $E_\pi r_j = H$. Thus, due to (3) and Taylor decomposition of the exponential function, MGF of $(Q_j - jH)/j$ is $(1 + O(j^{-2}))^j = \exp(-j)$.

The exponential Markov inequality applied twice for $(Q_j - jH)/j - \pm\varepsilon$ and the Bonferroni bound yield that the maximal absolute deviation on the interval $[\pm O(\log n)]$ has exponentially small probability.

It remains to apply the Borel-Cantelli lemma to finish the proof.

$S'_n = \sum_{i=1}^n m_{i,i-1}$ is a martingale. The Doob's maximal inequality for submartingale $\exp(tS_n)$ implies:

$$P_\pi(\max_{0<k\le n} S_k \ge \varepsilon)) \le E_\pi \exp(tS'_n)/\exp(t\varepsilon)$$

for every t. Now we find the appropriate t and ε.

By the MGF formula, section 5.1, $E_\pi \exp tS'_n = \pi^T G^n \mathbf{1}$, where $G_{ij} = P_{ij} \exp t[l_{ij} - \sum_{j=1}^B] P_{ij} l_{ij}]$, $1 \le i, j \le B$.

The mean $E_\pi m_{i,i-1} \equiv 0 = E_\pi S'_n$. Let us optimize an **exponential bound** for the maximal deviations of $\max_{0<k\le n} S'_k$ from its mean.

We have: $H_n(t) = n^{-1} \log E_\pi \exp tS'_n := \log R_n(t) \to \log R(t) := H(t)$

Bound optimization over parameter t.

Introduce $\sup(st - H(t))$, $\bar{L}(t) = \limsup_{\delta\to 0}[(L(t-\delta))(\mathbf{I}_{t>0}) + (L(t+\delta))(\mathbf{I}_{t<0})]$. Then under ET, it holds:

$$\limsup_{n\to\infty} n^{-1} \log P_\pi(\max_{0<k\le n} S'_k \ge \varepsilon)) \le -\bar{L}(\varepsilon).$$

The convex smooth Legendre transform of function $H(\cdot)$ is semi-continuous and positive for sufficiently small ε if $\sigma > 0$.

The above inequality and the **same inequality for** $-m_n$ imply inequality for the **absolute deviation**.

Proof is obtained via the Doob maximal martingale deviations and standard optimization in **exponential Markov inequality**, see e.g. [32] or [13], p. 410.

We choose $\varepsilon = k \log n$ in application to the **online Change Point detection** in section 9.2-3. It follows from the preceding development that the maximal absolute deviation of $m(\cdot)$ on an interval of length $k \log n$ is $O(n^{-q})$, $q = \bar{L}/k$, k is chosen to guarantee only finite number of violations of the absolute deviation bound according to the Borel-Cantelli lemma.

The functional CLT (see e.g. [3], theorem 2.11) states: if the conditional mean squared increments of square-integrable martingale satisfying Lindeberg condition converge to a const, then $m(\cdot)$ weakly converges to a Brownian motion under appropriate normalization. Conditions above are easily verifiable.

3. THE LOCAL ASYMPTOTIC NORMALITY OF SCOT

Given the context tree, denote the set of SCOT root-prediction probabilities satisfying natural normalization conditions by $\{\theta\}$. Their cardinality is $B \times B \leq B^2$ with B normalization conditions. We prove that the corresponding family of probability distributions id regular in the LAN-sense.

The principal role in the LAN proof is played by the multivariate AN of log-likelihood ratio as an example of multivariate ATF function (see section 4). For simplicity we assume that all entries of P_θ are positive.

The Local Asymptotic Normality (or simply LAN) introduced in Le Cam (1960) is the following decomposition of the local log-likelihood ratio

$$r_n(\mathbf{u}) = \ln[P_{\theta+n^{-1/2}\mathbf{u}}((X_0^n))/P_\theta((X_0^n))], \mathbf{u} \in \mathbf{R}^B = r(\mathbf{u}) + \psi_n(\mathbf{u}),$$
$$r(\mathbf{u}) = u^T\lambda - (1/2)\mathbf{u}^T J\mathbf{u},$$

where

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

is the limiting covariance matrix of $r_n(\cdot)$'s AN approximation assumed invertible, $J\mathbf{1} = \mathbf{0}_B, J^{-1}$ is called the Fisher information matrix.

and $\psi_n(\mathbf{u})$ converges in $P_\pi(X_0^n)$- probability to zero.

Proof. Applying the Taylor expansion of the second order

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

where

$$I = \mathbf{u}^T \partial r(\theta),$$
$$II = -\partial^2 r_n(\theta)\mathbf{u}^T J\mathbf{u}^T,$$
$$III = -\mathbf{u}^T J\mathbf{u}^T \partial r(\theta)^T J\partial r(\theta).$$

Now, $In^{-1/2} \to u^T\lambda$ weakly by CLT, section 5.2, $II/n \to 0$ by LLN, section 5.3, since $E\partial^2 r_n(\theta) = 0$, and $(1/2)III/n \to -(1/2)\mathbf{u}^T J\mathbf{u}$ again by LLN, section 5.3.

This expansion for a univariate parameter θ is proved in [31] referring to a much more involved exposition in [27] for the AN proof of the log-likelihood ratio 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. Namely, for all $K > 0, a > 0$

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

4. THE LOCAL ASYMPTOTIC MINIMAXITY OF THE LIKELIHOOD-RATIO-LIKE TESTS

We introduce the *Local Asymptotic Minimaxity* (LAM) and the Locally Asymptotically Most Power (LAMP) of the likelihood based inference and of its certain approximations. It is implied by the LAN condition outlined in section 6. Informally, the LAM in parameter estimation means that the deviation of the estimate from the true parameter value θ^* is asymptotically as minimal as possible in the local minimax sense.

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 $\delta > 0$:

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

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 [27,31]. [27] 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 ([7,27]) proves that the AN of a statistic under the null hypothesis implies its AN under the alternative distribution provided contiguity and the LAN condition.

5. 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 [27].

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 of the test satisfies $\beta_{n,\mathbf{u}} = P_{n,\mathbf{u}}(r_{n,\mathbf{u}} > C_{n,\alpha})$ 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.

5.1. Homogeneity testing

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 [21].

The first stage is estimation of the SCOT model of the string T following the algorithm in [19]. We refer to this publications for the details.

We assume

1. The T's and Q's are well-approximated by a sparse SCOT with Perfect Memory and

2. the LAN condition is fulfilled for the equivalent 1-MC over the space of contexts.

W cut the query string into K slices of the same length. Then, using the SCOT model of T we find the log-likelihoods $L_Q(k)$ of query slices Q_k and of strings S_k simulated from the training distribution of the same size as $Q_k, k = 1, \dots, K$, (for constructing simulated strings, see e.g. algorithm in [19]).

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 log-likelihood 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.

5.2. Comparison with GARCH on Apple log-returns

The first data set we use is the discretized in 27 bins daily log-return data of Apple Inc. starting from January 2, 2009

By observation, we pick the volatile region (the first 450 days returns) and the quiet region (the 500th to 600th days returns) to make a comparison. We first fit the data with the GARCH(1,1) modeled using the MATLAB(R2011a) GARCH toolbox. The P-values obtained are $P_1 = 0.0311$ and $P_2 = 0.0897$.

We apply SCOTlr to the same data. The homogeneity t-test between 1–450 and 500–600 (quiet and volatile regions) trained on 1–450 shows that the t-value is -16.02058. Thus, the P-value $P < 0.00001$. This P-value by SCOT is dramatically smaller than the Z-score by GARCH.

6. OFFLINE FITTING MARKOV SWITCHING MODEL

For notation simplicity we start with **two-state (± 1) HMM**.

Introduce stationary log-likelihoods $l_i = \log P(x_i|x_{i-1})$ and entropy of SCOT((± 1)): $h((\pm 1)) = \lim_{M\to\infty} M^{-1} E(\sum_1^M l_i)$.

6.1. HMM-SCOT model fitting

As explained before, the SCOT emissions can be reduced to a MC on the alphabet $\mathcal{B}$ of SCOT contexts.

We assume that diagonal elements of all Q hidden HMM X_n transition probability matrices are $a_i(l) = (1 - (c_i l)^{-1}) > 0, i = 1, \dots, Q, l \to \infty$ and emission distributions are all different. Emission SCOT sequences are assumed transformable to ergodic MC $y_i, \tau_j < i < \tau_{j+1}$ over the same alphabet $\mathcal{B}$ switching from a regime to alternative one at random CP time moments $\tau_j, j = 1, \dots, Q; \tau_0 = 0$.

We estimate both emission regimes, all CPs and HMM parameters. Thus, the HMM jumps to an alternative state after spending asymptotically exponentially distributed time with mean $c_i l$ in each state.

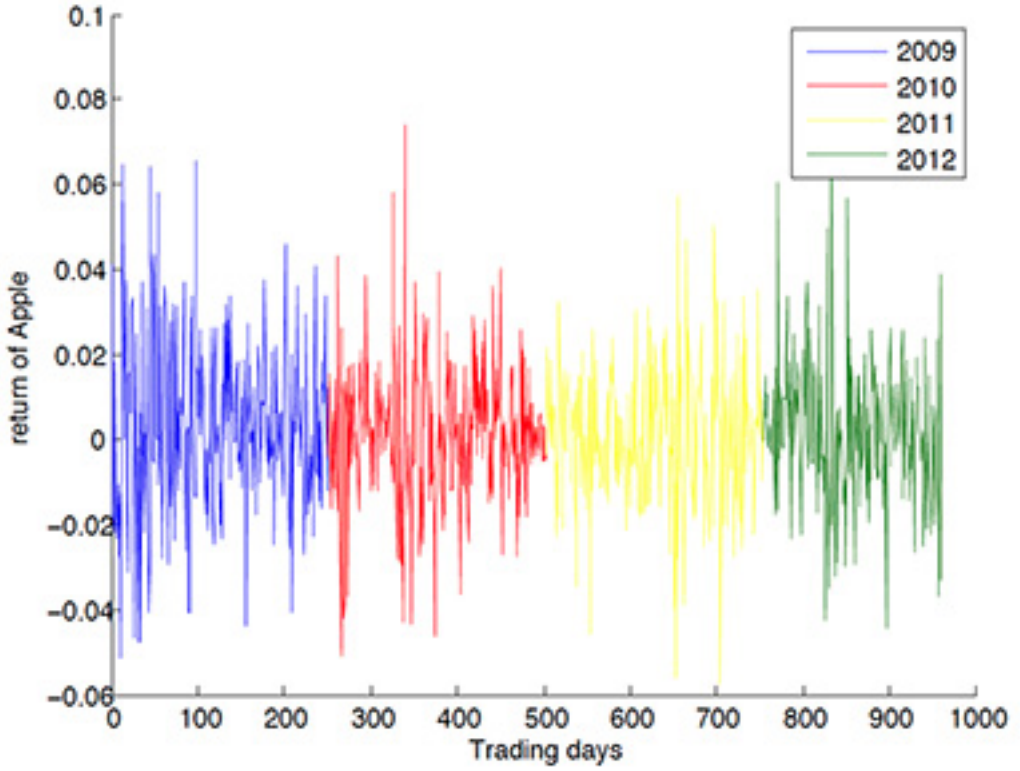

Fig. 2. Apple log-returns.

We modify the two settings of Change Point (CP) detection of IID sequences $m_i + \mu_j$ with changing mean $\mu_j, j = \pm 1$ in [18]. In their *offline* method, the quadratic risk of ML as a CP estimate does not exceed $O(1)$ as the sample size $n \to \infty$ due to certain exponential bounds. Their *online* method uses CP detection which is a point in the first N-interval, $N = O(\log n)$ such that the maximal absolute deviation of m_i deviation in this interval becomes comparable with the absolute change of mean μ.

Our plan for online CP detection is to replace their IID-based inequalities with the Doob-Meyer decomposition-based bounds for the maximal absolute deviation of both the martingale and compensator parts **separately**. We use the exponential bounds obtained in section 5.4. By implementing this program, we get the same order of quadratic risk as derived in [18] for IID case with changing mean.

Remark. A rougher estimates of the quadratic risk in online preliminary CP-detection can be obtained by a simpler quadratic maximum martingale deviation bound guarantying a larger window of order $O(\sqrt{n})$. It would require much larger risk and sample size, thus it is omitted.

Our algorithm uses repeatedly training (section 12) the SCOT emissions in all regimes and homogeneity testing of section 8.1.

To simplify notation, we first present our HMM-MC model with the two-state HMM (Q=2).

6.2. Algorithm road map, two-state HMM

For notation simplicity we start with the two-state HMM ($A = 2, b = \pm 1$). Our algorithm includes:

1. Online estimation of the $B \times B$-transition matrix P_{-1} of 1-MC equivalent to SCOT-emission-regime, see Appendices 1-2.

2. After some initial period of online continuous improving the P_{-1} estimate, start online detection of the first CP τ_1 in sequences of windows of size $O(logl)$ and find a preliminary first CP estimate. We show that its StD is $O(\log l)$.
3. The emissions from the time interval exceeding the CP estimate in $(l\varepsilon)^a, a > 0$ are used for online continuous estimation of the alternative regime P_1 and next CP τ_2.
4. Using $P_j, j = 0, 1$, estimates we update preliminary CP online estimates with the offline MLE. The offline MLE has StD=$O(1)$.
   Using the approach from the preceding item, we recurrently find estimates of all subsequent CPs and improve estimates of regimes $P_j, j = \pm 1$.
5. Using all CP estimates, we get MLE of the HMM parameters.

6.3. Online CP detection

Given a 1-MC transition matrix in region 1, we carry on the CP online detection between regimes 1 and -1.

We choose such a window size that within window

$$P(\max|m(k)| > 0.1(\Delta h)\sum c(k)) < l^{-3}, \tag{6}$$

where Δh is the entropy difference between the current and alternative regime. A lower bound for Δh is used if Δh is unknown. This window size is evaluated using the maximal inequality for absolute value of martingale $|m(\cdot)|$, section 4.2. Our CP-detector is the first window when (5) occurs.

The Borel-Cantelli lemma implies that only finite number of events (5) occur under $l \to \infty$. As follows from section , the window size is $O(|\log l|)$ which implies the same order of the standard deviation of our CP detector.

6.4. Offline CP detection

Our offline segmentation stage estimates time regions with constant HMM states using homogeneity test for SCOT emission strings and a preliminary online segmentation. This is made fast recurrently in parallel on a cluster of computers.

The offline CP detection follows after the online CP estimate is obtained. It starts with the **SCOT training of the string after small delay of length** $O(\log n)$. The homogeneity test verifies significance of new regime distinction from the previous one. The offline CP update of the preliminary online CP estimate is the location of the maximum of the log-likelihood function.

For simplicity of notation assume that the initial regime is P_{-1} and the **time of the first CP is** 0. Introduce a surrogate 'log-likelihood function' L_z under 'possibly false CP' at time z and show that $\max L_z$ is attained at a point $z_* : E(z_*)^2 = O(1)$.

Family $L_z = I + II + III$ is irregular and methods of section 6 are inapplicable. First suppose $z < 0$ and $[\pm n]$ is included into exactly two regimes ± 1; 0 belongs to the online CP interval estimate.

We have $I = \sum_{-n}^{z} l(X(k), X(k-1))$,

$II = \sum_{z}^{0} \lambda(X(k), X(k-1))$,

$III = \sum_{0}^{n} \lambda(X(k), X(k-1))$,

where $Q_{X(k),X(k-1)}, \lambda(X(k), X(k-1))$ are regime $(+1)$ transition probabilities and their logarithm.

Every summand in II has mean $E_P\lambda = E_P(l) - \delta_{-1}$. Thus $E_P(L_0 - L_z) \geq z\delta_{-1}$.

The case $z > 0$ is dealt with quite similarly resulting in equality $E_Q(L_0 - L_z) \geq z\delta_{+1}$. Both cross-entropies $\delta_{\pm 1}$ are positive.

The offline interval CP estimate can be updated in many ways including bification for iterative numerical finding z_*.

It remains to bound its quadratic risk from above. The lower bound of the same order of magnitude follows from the IID case with changing mean in [18].

Lemma 5.4 implies that the compensators for L_z are maximal at a point $O(1)$. Convergence of the normalized maximal absolute difference between the compensator and its mean H_n to 0 follows from Lemma 5.4.

The functional CLT for the martingale component of log-likelihood [3], theorem 2.11, states that the normalized martingale sequence converges weakly to the Brownian motion. Our exponential bounds show that the maximal deviation converges to 0 also in L^2.

The normalized left log-likelihood over negative times $[-n, 0]$ was proved in section 5.2 to be asymptotically Normal with positive left slope and variance σ_-^2. Similarly, the normalized right log-likelihood of the **reverted MC** (which is also ergodic) over positive times has negative slope at 0 and variance σ_+^2. Thus, the quadratic risk of z_* is $O(1)$ in the weak convergence sense. The mean square convergence can be justified in a standard way.

The SCOT is proved to be LAN which implies that the same orders of quadratic risk remain valid when using the estimated SCOT parameters during search for CP.

6.5. HMM with finite number of states

Training SCOT for the general m states slow HMM model such that all time means spent in states before jump are proportional to large parameter n. Main steps of training are similar to the two HMM state case. Online change point detection is used before every jump to unknown state.

Main steps of training are similar to the two-states HMM case. Online change point detection is used before every jump to unknown state. It is followed by the SCOT training of the string using some delay after jump, where homogeneity is verified by homogeneity test and by the subsequent Maximum likelihood offline change point update of the preliminary CP online estimate as above.

After all change points are safely estimated, parameters of HMM are ML-estimated based on their multivariate statistics.

6.6. HMM parameters estimation

.

If it is only known that all mean times spent in every state before jump are proportional to large parameter l, we can estimate all HMM transition probabilities after detecting all jump times. The marginal HMM distribution is estimated via maximum likelihood applied to the joint CP estimates using obvious frequencies. Namely, denoting n_{ij} = number of times i is followed by $j, j = \pm 1$, under the stationary initial distribution, the log-likelihood is (see e.g. [2])

$$l(p) = \sum_{ij} n_{ij} \log p_{ij}, \sum_j p_{ij} \equiv 1,$$

which yields the ML estimates

$$\hat{p}_i j = n_{ij} / \sum_j n_{ij}.$$

Thus, the average of empirical mean times before estimated jump from i to any $j \neq i$, serves as an estimate of mean time spent in i, while transition probability from i to any $j \neq i$ is estimated via the last formula or simply as the frequency of jumps from i to j out of all visits to i.

6.7. Confidence band for HMM parameters

The above estimation of the HMM parameters p_{ij} is a regular statistical problem with non-degenerate Fisher Information matrix $I(p) = E_\pi \partial l \partial l^T / n$, the offline CP estimates are independent 'observations' with square risks of order $O(1)$ not affecting asymptotics of the quadratic risk of ML-estimates. Thus asymptotically,

$$Cov(\hat{p}) = [I(p)n]^{-1}$$

.

For large state space, iterative methods via Fisher scores are available.

7. DISCUSSION AND ACKNOWLEDGMENTS

Our display of 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, exponential bound for the martingale part of log-likelihood and its application for the online and offline CP detection seem new.

A challenging open problem is to prove accurate asymptotic results for MC alphabet rising simultaneously with the sample size.

Appendix 3 reviews results of [35], several revised parts of [20] are used elsewhere in the text including a simulation prepared by P. Grosu. The author is deeply grateful to them for the long collaboration and help.

REFERENCES

1. Aminikhanghahi S. and Cook D. A Survey of Methods for Time Series Change Point Detection, *Knowl Inf. Syst.*, 2017, vol. 51, no. 2, pp. 339–367.
2. Billingsley P. *Statistical inference for Markov chains*, University of Chicago Press, 1961.
3. Biscup M. Recent progress on the Random Conductance Model, *Probability Surveys*, **8**, 294–373, 2011.
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. Brodsky B. E. and Darkhovsky B. S. *Nonparametric methods in change-point problems*, Kluwer, Dodrecht, 1993.
7. Chibisov D. M. Lectures on the asymptotic theory of rank tests, *Lecture Notes NOTs 14. M.: Matematicheskiy 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. Cappe O., Moulines E., Rydyen T. *Inference in Hidden Markov models*, Springer, 2005.
10. Dumont T. Context Tree estimation in Variable Length Hidden Markov Models. *IEEE Trans. Inform. Theory*, 2014.
11. Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis*, Cambridge University Press, 1998.
12. Feller W. *An introduction to Probability theory and its applications, volume 1, third edition*, Wiley, N. Y., 1967.
13. Gallager R. *Stochastic processes: Theory for applications*, Cambridge Uni., 2013.

14. 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., 2008, pp. 117–134.

15. Grinstead C. and Snell J. *Introduction to Probability*, AMS, 2006.

16. Hamiltons J. *Regime switching models*, The New Palgrave Dictionary of Economics. Second Edition, Eds. Steven N. Durlauf and Lawrence E. Blume, Palgrave Macmillan, 2008.

17. Kharin Yu. and Maltsau M. Markov Chain of Conditional Order: Properties and Statistical Analysis, Austrian Journal of Statistics, **43/3-4**, 2014, 205–216.

18. Korostelev A. and Korosteleva O. *Mathematical Statistics: Asymptotic Minimax Theory*, AMS, Providence, R.I., 2011.

19. Mächler M. and Bühlmann P. Variable Length Markov Chains: methodology, computing, and software, *Journal of Computational and Graphical Statistics*, 2004, vol. 13, no. 2, pp. 435–455.

20. Malyutov M. and Grosu P. SCOT approximation, modeling and training, *Proceedings of Machine Learning Research*, 2017, vol. 60, pp. 241–265.

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

22. Markov A.A. Extension of the limit theorems of probability theory to a sum of variables connected in a chain (1906, in Russian), reprinted in *Appendix B of: R. Howard. Dynamic Probabilistic Systems, volume 1: Markov Chains.* John Wiley and Sons, 1971.

23. Malyutov M. Retrospective Training Slow HMM-SCOT Emissions Model, *Information Processes*, 2017, vol. 17, no. 3, pp. 199–205.

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

25. Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of IEEE*, , 1989, vol. 77, no. 2, 257–286.

26. Rissanen J. A universal data compression system. *IEEE Trans. Inform. Theory*, 1983, vol. 29, no. 5, pp. 656–664.

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

28. Ryabko B., Astola J. and Malyutov M. *Compression-Based Methods of Prediction and Statistical analysis of Time Series: Theory and Appllications*, Springer International, 2016.

29. Shiryaev A. N. *Optimal stopping rules*, Applications of Mathematics, 1978, vol. 8, Springer, New York.

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

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

32. Veretennikov A. Yu. Ergodic Markov Chains and Poisson equations (lecture notes),*Arxiv:1610.09661v3 [math.PR]*, 2017.

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

34. Yoon B.J. Hidden Markov Models and their Applications in Biological Sequence Analysis, *Current Genomics*, 2009, vol. 10, no. 9, pp. 402–415.

35. Zhang T. Perfect Memory Context Trees in time series modeling, *Information Processes*, 2017, vol. 17, no. 1, pp. 70–81.

APPENDIX 1. PARALLEL SCOT TRAINING

This section outlines a novel parallel implementation of the algorithm similar to 'Context' which is created for fast processing more complex data sets including those with larger alphabet sizes.

The ESI-based criterion usually stops back-processing of the training string long before the chosen horizon. All directions backwards from the root are processed in parallel making the algorithm much faster. written using the Python programming language – builds the stochastic trees starting from stage 1 and proceeding to the horizon stage of interest. Potential contexts having an ESI value smaller than $\epsilon > 0$ become contexts, and would be omitted from processing in the following stages. Another improvement in parallelism is processing of a potential context by hashing, and determining if it should be processed on one node of many by taking the modulo of the hash with the total number of compute nodes. The assumption here is that there we have many (hundreds) of compute nodes available to process a corpus into a SCOT.

ESI evaluation for possible context

The Empirical Shannon Information (ESI) is an approximation to the log-likelihood ratio statistic for the root independence from all more remote symbols than a putative context..

Given a context s, let $N(s)$ be the count of s in the source. Define a function $ESI(s)$ as follows:

$$ESI(s) = \sum_{i \in A} \sum_{j \in A} N(i.s.j) \cdot log_2 \left(\frac{N(i.s.j) \cdot N(s)}{N(i.s) \cdot N(s.j)} \right).$$

Deciding about contexts

Using a fixed maximum context length (horizon) h and a threshold $\epsilon > 0$, we define a *context* over source as follows:
It is decided that string $x_1, x_2, \ldots, x_t,\ t \leq h$ is a context, if and only if:

(a) For any i, $i = 2, \ldots, t$ such that $ESI(x_1, x_2, \ldots, x_t) > \epsilon$,
(b) $0 < ESI(x_1, x_2, \ldots, x_t) \leq \epsilon$.

Building SCOT

Using our criterion on *contexts*, we check all the messages coming from a source and build a SCOT such that each *context* is a path starting from a leaf and ending at a son of the root. A SCOT is built in a step-wise manner starting with depth 1 and ending at the desired *context* length.

The stochastic component of SCOT - prediction distribution of symbols in the root - is to be specified at every leaf. This is performed by the following equation, where s is a leaf:

$$P(i|s) = N(s.i)/N(s), \text{where } i \in A.$$

These algorithms to build up the contexts are always started with Stage 1. At the end of each stage the contexts are collected from all the nodes and stored into one file, which will be used as an input for the following stage. This writing of the files by each node is equivalent to the Map phase of the MapReduce algorithm, while the collection is the Reduce phase of it. In order to build the final SCOT file to a specific stage, all the contexts to the desired stage are combined and the horizon is consolidated with all the non-contexts of only the specific stage.

In making the SCOT construction parallel, the choice for parallelism are substrings which will be tested for being contexts via the $RissanenESI$ function. Such potential contexts would require the

prefix and suffix counts to be populated. To parallelize the processing of contexts, a hash function was implemented to construct a numerical value out of the characters forming the context string, which subsequently is modulo with the number of nodes (*total_nodes* variable). This way each compute node will process a portion of the potential contexts, which can be many in latter stages of a SCOT. Our hash algorithm for strings is standard.

The large prime number 99971 is used in how the data is actually stored to load-balance the retrieval of sequence counts. Each sequence thus would be stored in a list (chained) in the bucket denoted by the modulo – because of possible collisions.

An annotated computer code with a detailed description of the algorithm is in [20], section 8.

APPENDIX 2. SIMULATION OF CP DETECTION (J. FENG)

We consider a HMM with states 0 and 1: starting with HMM state 0, we use the SCOT model 2 from [28], the transition probability to state 1 is 0.001. Increasing SCOT under state 0: $z_k = k$ (that means the probability of changing from state k-1 to state k is 1, starting from the initial state $a_0 = 0$) Decreasing SCOT under HMM state 1: the probability of changing from state k to k-1 is 1. Uniformly distributed random numbers u_k with density f(x)=1, and range [0,1].

Model 2 (i), [28]

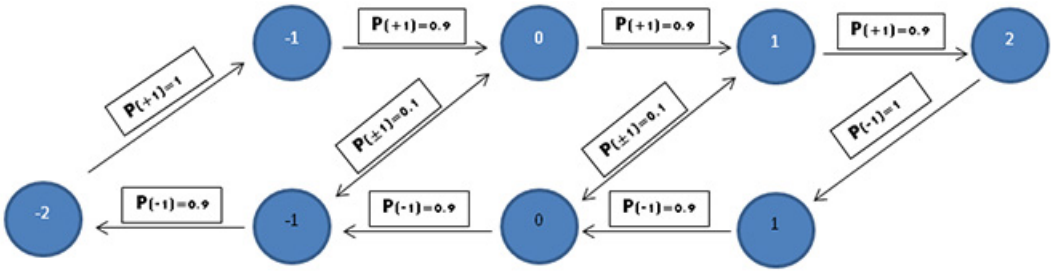

Fig. 3. The scheme of Model 2(i) is displayed

This model does not lead to undesirable periodicity of the MC over contexts because our Change Point precedes the first hit of the boundary.

$X_n = X_{n-1} + 1$ with probability 0.9, if $X_{n-1} = X_{n-2} + 1$ and $X_{n-1} \neq \pm l$;

X_n is defined anti-symmetrically to the above, if $X_{n-1} = X_{n-2} - 1$ and $X_{n-1} \neq \pm l$.

Here is the print out for the observations $y_k = u_k * 0.99 + z_k * 0.01$:

In simulation of the online change point detection the online CP is found by doing the following:

1. Under the assumption that the sequence stays in HMM state 0 for the whole time, we compute the log likelihood of observations y_k.

2. Compute function F1(k)= sum of the log likelihood up to time k

3. The above function F1(k) should be increasing.

4. Under the assumption that sequence stay in state 1 all the time, we compute the log likelihood of observations y_k.

5. Compute function F2(k)= sun of the log likelihood from time k to the end time using the inverse Markov chain. 6. The above function F2(k) should be increasing as time decrease.

7. The point where F2(k) and F1(k) intersect is the CP ML-estimate.

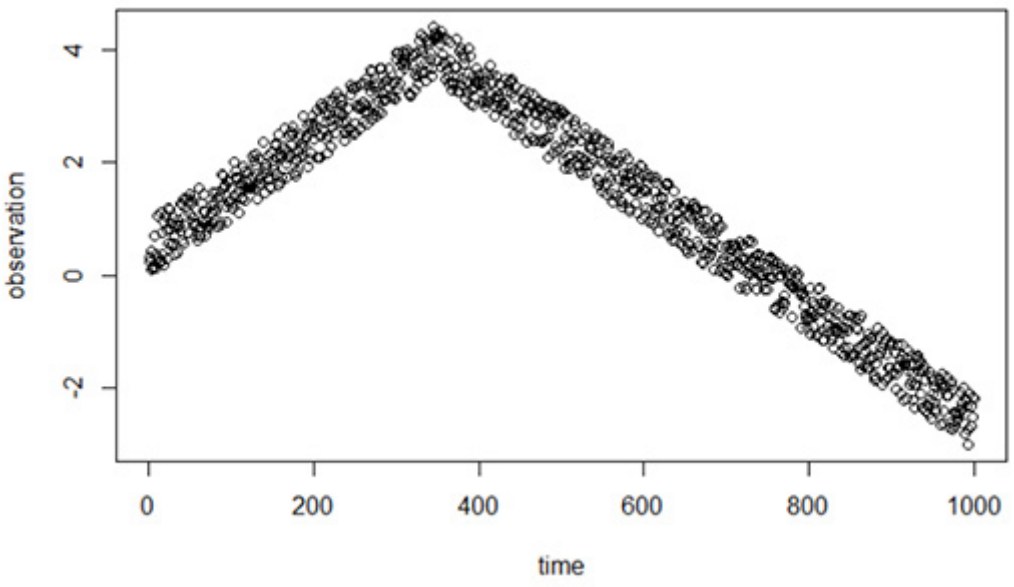

Fig. 4. Simulated HMM-SCOT model.

Case 1:

Set up:

i) Let U be the random variable with discrete uniform distribution.
i.e. P(U=j/10)=1/100 for j=1,2,...,100.

ii) Let s_t indicate the state of Hidden Markov Model at time t (with only two state 0, 1).

The transition probability of this Hidden Markov Model is:

$P(s_1 = 0) = 1, P(s_t = 1|s_{t-1} = 0) = 0.001, P(s_t = 1|s_{t-1} = 1) = 1.$

So the Hidden Markov Model will start with state 0 and then has one change point after some time (the state changes from 0 to 1) and will never go back to state 0 again.

Denote CP as the least t such that $s_t = 1$.

iii) Define increasing SCOT under state 0 in Hidden Markov Model:

Let z_t act under the rule of increasing SCOT if $s_t = 0$, that is:

$P(z_1 = 1) = 1, P(z_t = k|z_{t-1} = k - 1) = 1.$

Then $P(z_t = t) = 1$.

iv) Define decreasing SCOT under state 1 in Hidden Markov Model:

Let z_t act under the rule of decreasing SCOT if $s_t = 1$, that is:
$P(z_t = k|z_{t-1} = k + 1) = 1.$

v) Summarizing ii) to iv), we have:

$z_t = 1$, if $t = 1$,

$z_t = t$, if $1 < t < CP$,

$z_t = 2CP - t - 2$, if $t \geq CP$.

vi) Define $y_t = 0.99U_t + (0.01)z_t$.

For simulation, we simulate y_t by simulating U_t , Hidden Markov Model and SCOTs. Then we want to use the Change point detection to find the change point by using the data we get for y_t s

and see how close the estimation is to the real change point CP. In this simulation, the sample size will be 1000 (i.e. $t = 1, ..., 1000$).

Online change point detection:

a) Calculating the log likelihood of each y_t:

The distribution of the CP:

$P(CP = i) = 0.99^{(i-2)} \times 0.01$ for $1 < i \leq t$,

$P(CP > t) = 0.99^t$.

The distribution of z_t given the value of CP:

$P(z_1 = |CP = i) = 1$,

$P(z_t = 2CP - t - 2|CP = i) = 1$ for $1 < i \leq t$,

$P(z_t = t|CP > t) = 1$ for $i > t$.

The probability of getting $y_1, y_2, ..., y_t$:

$$P(y_1, y_2, ..., y_t|CP = i) = \sum_{\substack{u_1,...,u_t \\ z_1,...,z_t}} P(0.99u_m + 0.01z_m = y_m \quad \forall m = 1, ..., t|CP = i)$$

$P(y_1, y_2, ..., y_t) =$

$\sum_{i=2}^{t} P(y_1, y_2, ..., y_t|CP = i)P(CP = i) + P(y_1, y_2, ..., y_t|CP > t)P(CP > t)$

Then the log-likelihood of $y_1, y_2, ..., y_t$ is $l_t = \log P(y_1, y_2, ..., y_t)$.

b) Average log-likelihood :

Choose the window size of 10 points, then the average log-likelihood from the window $y_t, \ldots, y_{(t+9)}$ is

$(\bar{l_t}) = 1/10 \sum_{k=0}^{9} l_{(t+k)}$.

c) Calculating the trend L_t of the data:

We use Least square estimate for model $l_k = ak + b$ with data $l_1, l_2, ..., l_t$ then $L_t = a$.

d) getting critical point:

Let C denote the critical point, by using the first hundred datas, we have first 100 $L_t's$.

Define $V(t) = |(\bar{l_t}) - Lt|$ Then let $C = 1.1 \max_{t \leq 100} V(t)$

e) Online change point estimate:

The estimator is the least t such that $|(\bar{l_t}) - Lt| > C$.

Simulation result

Picture 1:

True change point is 389. $C = 17.768$ The least t such that $V(t) > 17.768$ is 383

Offline change point detection:

we use maximum likelihood estimator as our estimation:

Define $L(\theta) = L(y_1, ..., y_t; \theta) = P(y_1, ..., y_t|CP = \theta)$ where $P(y_1, ..., y_t|CP = \theta)$ can be deduced from similar procedure introduced in online change point detection. Then the maximum likelihood estimator θ_0 is:

$\theta_0 = arg \max_{1 \leq \theta \geq 1000} L(\theta)$

Simulation result:

Picture 2:

True change point is 389. Offline change point estimater $\theta_0 = 389$

Case 1: Set up i) Let U be the random variable with discrete uniform distribution. i.e. P(U=j/10)=1/100 for j=1,2,...,100.

ii) Let s_t indicate the state of Hidden Markov Model at time t (with only two state 0, 1).

The transition probability of this Hidden Markov Model is:

$P(s_1 = 0) = 1, P(s_t = 1|s_{t-1} = 0) = 0.001, P(s_t = 1|s_{t-1} = 1) = 1.$

So the Hidden Markov Model will start with state 0 and then has one change point after some time (the state changes from 0 to 1) and will never go back to state 0 again.

Denote CP as the least t such that $s_t = 1$.

iii) Define increasing SCOT under state 0 in Hidden Markov Model:

Let z_t act under the rule of increasing SCOT if $s_t = 0$, that is:

$P(z_1 = 1) = 1, P(z_t = k|z_{t-1} = k - 1) = 1.$

Then $P(z_t = t) = 1$.

iv) Define decreasing SCOT under state 1 in Hidden Markov Model:

Let z_t act under the rule of decreasing SCOT if $s_t = 1$, that is:
$P(z_t = k|z_{t-1} = k + 1) = 1.$

v) Summarizing ii) to iv), we have:

$z_t = 1$, if $t = 1$,

$z_t = t$, if $1 < t < CP$,

$z_t = 2CP - t - 2$, if $t \geq CP$.

vi) Define $y_t = 0.99U_t + (0.01)z_t$.

For simulation, we simulate y_t by simulating U_t , Hidden Markov Model and SCOTs. Then we want to use the Change point detection to find the change point by using the data we get for y_t s and see how close the estimation is to the real change point CP.

Online change point detection:

a) Calculating the log likelihood of each y_t:

The distribution of the CP:

$P(CP = i) = 0.99^{(i-2)} \times 0.01$ for $1 < i \leq t$,

$P(CP > t) = 0.99^t$.

The distribution of z_t given the value of CP:

$P(z_1 = |CP = i) = 1$,

$P(z_t = 2CP - t - 2|CP = i) = 1$ for $1 < i \leq t$,

$P(z_t = t|CP > t) = 1$ for $i > t$.

Then: $P(z_t = x) = \sum_{i=2}^{\infty}[P(z_t = x|CP = i)P(CP = i)]$

$= \sum_{i=2}^{t}[P(z_t = x|CP = i)P(CP = i)] + P(z_t = x|CP > i)P(CP > i)$.

The probability of y_t equal to the actual observation c:

$P(y_t = c) = \sum_{j=1}^{100}[P(U = j/10)P(z_t = 100c - j/10 * 99)]$.

Then the log-likelihood of y_t is $l_t = \log P(y_t = c)$.

b) Average log-likelihood : Choose the window size of 10 points, then the average log-likelihood from the window $y_t, \ldots, y_{(t+9)}$ is

$(\bar{l_t}) = 1/10 \sum_{k=0}^{9} l_{(t+k)}$.

Calculating the trend L of the data: We use Least square estimate for model $y_t = at + b$ then $L = a$.

Change point estimate: Our CP estimate is the least t such that $|(\bar{l_t}) - kL| > 0.1$.