Информационные процессы, Том 17, № 3, 2017, стр. 199–205

Retrospective Training Slow HMM-SCOT Emissions Model

M. Malyutov*

Mathematics Dept., Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA
e-mail: m.malioutov@neu.edu, pgrosu@gmail.com
Received 15.09.2017

Abstract—The Baum–Welsh recurrent ML estimation of HMM parameters has been successfully applied to speech recognition. Its application to Genome modeling is questionable since assigning independence and equal probabilities to emissions from the same part of genome is a rough approximation. We develop a hybrid slow HMM switching model with SCOT emissions which might be a more realistic model for Genome, analysis of combined authorship of literary works, seismological data or financial time series with piecewise volatility. Our combined online and offline segmentation stage estimates time regions with constant HMM states using homogeneity test for SCOT emissions strings. This is made recurrently in parallel on a cluster of computers.

KEYWORDS: SCOT models, Baum-Welsh algorithm, HMM, regime switching model.

1. INTRODUCTION

The popularity of a compression tool – VLMC based on SCOT fitting has been increasing rapidly since a fitting algorithm was proved to be *consistent for stationary mixing sequences* in [13] and used for compression. A somewhat confusing name VLMC is also used for a video editor.

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. The parallel super-fast training of a **sparse** SCOT model and asymptotically optimal inference about the SCOT model including the nonparametric homogeneity test are described in [10].

Markov *regime switching* models remain enormously popular in speech recognition, economics, finance, etc. A concise review is in [6]. Nonparametric segmentation in switching models without probability assignment of jump moments is in [1] and many consequent papers by these authors. The Hidden Markov Model (HMM) is the simplest regime switching model with all regimes consisting of random variables called **emissions**. Emissions are independent (mutually and of HMM) with distribution depending on the current HMM state. The fast Baum-Welsh recurrent retrospective ML estimation of HMM parameters has been successfully applied to speech recognition [12]. Its application to Genome modeling [4, 17] is questionable since assigning independence and equal probabilities to emissions from the same HMM state (which is the same part of Genome in this model) is a rough simplification. Markov switching models generalize HMM by considering more complex **parametric** regimes admitting fast parameter estimation ([2, 6]).

We develop semi-parametric hybrid slow HMM with SCOT emissions (SCOT-HMM) which might be a more realistic model for Genome, analysis of combined authorship of literary works, seismological data or financial time series with piecewise volatility. We show in [10, 11] that regular stationary sequences are well-SCOT-approximated. Thus, SCOT-HMM models are close to nonparametric ones.

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

It is necessary to distinguish our HMM-SCOT model from a completely different VLMCHMM approach introduced in [3].

A more **formal SCOT description** follows.

An m-MC $\{X_n\}$ with a finite state space (alphabet) A can be regarded as 1-MC

$$\{Y_n = (X_n, X_{n+1}, \dots, X_{n+m-1})\}$$

with alphabet as the space of *m-grams* A^m. Namely:
$P(Y_{n+1}|Y_n) = P(X_{n+m}|Y_n)$, if $X_{n+1}, \dots, X_{n+m-1}$ coincide in both sides, and 0 otherwise.

This induced MC on A^m is not necessarily ergodic for ergodic m-MC. A simple counterexample follows.

1.1. Counterexample

Consider binary 2-MC with alphabet $\{0, 1\}$ and transition probability 1/2 from $\{0, 1\}$ or $\{1, 1\}$ to $\{0\}$ and $\{1\}$, transition probability 1 from $\{0, 0\}$ and $\{1, 0\}$ to $\{1\}$. It is ergodic, but the induced MC on 2-grams has transient state $\{0, 0\}$.

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.

A Sparse SCOT over some alphabet A is a very special case of m-MC, where m is the **maximal length of contexts**. For a given string $x_0^- := (x_{-m}, \dots, x_{-1}, x_0)$, the *context to a current state* x_0 *(***root***) given preceding m-gram* is

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

the end part of the preceding m-gram of **minimal length** such that the conditional probability

$$P\left(x_0|x_{-1}^{-r}\right) \equiv P\left(x_0|x_{-1}^{-k}\right), \forall r > k; k = \left|C\left(x_0^-\right)\right| \tag{2}$$

is called the *length of context* $C\left(x_0^-\right)$.

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 non-terminal node of the context tree.

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. Averaging context lengths over their SCOT stationary distribution gives a sparsity indicator—the mean prediction length (MPL). SCOT is called sparse if MPL=$o(m)$. The median, or another quantile collection, or other functions of $\mathcal{M}$ over the stationary distribution can be also used for defining sparsity. Notice that for $n < m$, n-step prediction can use states on distance from m to 1 from the root in contrast to prediction in traditional theory of stationary processes.

The SCOT stationary distribution is evaluated in [14] either analytically or iteratively using reduction of SCOT to a 1-MC on the space of contexts [14, 18].

Assumptions. A sparse SCOT is always assumed to be **ergodic** and its **maximal context length** M **(horizon) is fixed**, the **number** B **of HMM states is fixed** while the sample size N is growing.

If 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 [5, 7, 10, 13].

For completeness, we display in section 2 a sketch of the Context algorithm consistency from [5] admitting horizon $O(\log N)$ and the sample size $N \to \infty$. A popular software [7] implementing the Context algorithm of [13] assumes a fixed horizon as $N \to \infty$. We sketch simplifications in proving consistency under this assumption.

2. CONSISTENCY OF SCOT TRAINING

$P(y, x_M, ..., x_1)$ is the joint empirical distribution of $(y = x_0, x_M, ..., x_1)$.

For a node $A = x_i, ..., x_1, P(y|zA) = P(y|A)/P(y|zA) = P(y, zA)/\sum_{b\in A} P(b, zA)$ is the conditional empirical distribution of zA given A, where zA denotes the string $(z, x_i, ..., x_1)$.

The Empirical Shannon Information (ESI)

$$\mathcal{I}_A = \sum_{y,z} P(y|zA) \log[P(y|zA)/P(y|A)], \tag{3}$$

$$T(A) = N(A)\mathcal{I}_A, \tag{4}$$

where $N(A) = \sharp$ of node A in the string.

Test $\mathcal{T}_\varepsilon$ of [13] chooses as contexts such nodes A that $T(A) < \varepsilon$.

Consistency proofs of SCOT contexts estimation of [13] and his followers has admitted possibly growing as $\log N$ maximal context size (horizon) for sample size $N \to \infty$.

To prove that the estimate $\hat{l}$ of the *length of context* $l(A) = l\left(C(x_0^-)\right)$ is the true one, they upper bound the probability of the opposite event by

$$P(\hat{l}) \neq l(A)|N(A) > C_1 N/\sqrt{\log N})P(N(A) > C_1 N/\sqrt{\log N}) + P(\cup_{\alpha\in A} N(\alpha \leq C_1 N/\sqrt{\log N})).$$

Rissanen proves that the first term is bounded by $C_2 \log N \exp(-C_3\sqrt{N})$.

The second term goes to 0 due to the ergodicity of the time series concluding the proof.

2.1. Consistency under Fixed Horizon

To simplify the proof and sketch the rate of consistency and conditional accuracy of prediction distributions assignment in the contexts, we use more practical assumption of fixed maximal context size $M = const$ as $N \to \infty$. We assume that the minimal cross entropy between the prediction distributions at nodes of the memory tree immediately preceding the context A or following context A exceed $\varepsilon + \delta, \delta > 0$. Then, fulfilling inequality $T(A) < \varepsilon$ is a large deviation with exponentially small in N probability. The Bonferroni bound means multiplication by a fixed under $N \to \infty$ multiplier and does not affect the exponential decay of the error probability. Conditional to the correct decision about a context, the prediction distribution in the root given the context has a degenerate multivariate normal distribution estimated by $P(y, A)/\sum_{b\in A} P(b, A)$.

Remark. Assuming a finite horizon M can be interpreted as follows: we replace the original m-MC with another one. Its conditional probabilities are replaced with averages of the original ones over their stationary distributions with respect to the tails of length exceeding M. Due to exponential memory loss of regular stationary processes, this approximation seems appropriate.

3. PRELUDE: DETERMINISTIC HMM

To make our ideas more transparent, we first display them in section 3 for toy examples of deterministic HMM. To warm up, consider first a much simpler than HMM case of B alternating

intervals of length L where states are respectively ± 1 and the beginning of intervals and maximal context sizes are known. The minimal absolute entropy difference between SCOT stationary distributions is assumed exceeding $\lambda > 0$.

Combining all intervals corresponding to the same states into long strings $T_i, i = \pm 1$, we can train each of their SCOT model according to [10].

Remark. To initialize SCOT sub-string, we use in this and following sections a uniformly distributed sample of the maximal length of their contexts which is presumed to be known beforehand. The cardinality of the sample is reasonably small since SCOT is assumed sparse. For L large enough, this initialization does not affect consistency of SCOT parameters estimation.

3.1. Quasi HMM

i. Known SCOT.

We consider first the same model with B = 1, the beginning of the first interval is **shifted right** to unknown value Δ_0 which is uniformly distributed over $[0, L]$ and the second interval is shifted accordingly, its end fills the interval $[0, \Delta_0]$. Thus, our assignment is on *circumference* $[0, 2L]$. The *known* SCOT(± 1) structures are assigned respectively to $[\Delta_0, L + \Delta_0]$ and the remainder.

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

To reduce analysis to the previous trivial case, we estimate Δ_0. For each $0 < \Delta < L$ we compute the normalized log-likelihood difference R/L of part $[\Delta, \Delta + L]$ vs the opposite part $[\Delta + L, \Delta + 2L]$ on circumference $[0, 2L]$. Finally, $\hat{\Delta}$ which maximizes over Δ the maximal over ± 1 value of $|R(\Delta)|$ is an estimate for the change point Δ_0. If $\hat{\Delta} = 0$ or L, then each of them is an equivalent estimate. For large L and $\Delta = \Delta_0, E(R/L)$ is the entropy difference $H_{+1} - H_{-1} > 0$ between SCOT(± 1). Otherwise, it is the entropy difference between their mixtures.

Proposition. $\hat{\Delta}$ is consistent with variance $O(L^{-1})$, if $|E(R(\Delta_0))/L| > 0$. Graphically, $E|R(\Delta)|$ is represented by two intervals with slopes $\pm(H_{+1} - H_{-1})$ which attain maximal values $H_{+1} > H_{-1}$ around the point $\hat{\Delta}$ of their local maximum.

Remark. If instead of condition $|E(R(\Delta_0))/L| > 0$ we use contiguous alternative SCOT distributions (which might be appropriate for literary texts of two authors), then a slightly modified procedure should be used, see [8, 10].

Proof of Proposition. Unless $\Delta = \Delta_0$, the log-likelihoods are mixtures and R/L are weighted differences of the SCOT(± 1) log-likelihoods. Due to the ergodic theorem ([10]), their averages converge to their expectations which implies consistency of $\hat{\Delta}$. The maximal values of $E|R(\Delta)|$ around Δ_0 are $H_{+1} > H_{-1}$, the first right and left differentials are proportional to $\pm(H_{+1} - H_{-1})$. Therefore, the asymptotic variances of the limiting one-sided Normal distributions are $Var_{\pm 1} l L^{-1}(H_{+1} - H_{-1})^{-2}$.

Thus, we estimate the change point Δ_0 with standard deviation $O(L^{-1})$.

ii. Unknown SCOT.

If SCOT(± 1) are unknown, we divide $0, L$ into, say, 5 equidistant slices and choose two adjacent slices with minimal t-value of the homogeneity test [10] which we use for training SCOT denoted as SCOT$(+1)$. We test homogeneity of the mirror slices in $[L, 2L]$ and after homogeneity confirmation train SCOT denoted as SCOT(-1). If homogeneity of mirror slices is not confirmed, we repeat the whole procedure anew until success.

Next, we estimate the change point Δ_0 similarly to item i. If $\hat{\Delta}$ turns out to belong to adjacent homogeneous slices we started SCOT training with, we repeat the whole procedure anew. Another approach is to apply online-offline sequence of change-point detections introduced in section 4.iii. which seems inferior for this simple case.

4. HMM WITH TWO STATES

i. Known HMM and unknown SCOT. For HMM with two states ± 1 and positive entropy difference between SCOT(± 1) distributions like in the preceding Proposition, we must estimate SCOT structures during HMM state constancy. For simplicity, we assume the transition probability matrix of HMM to be the $\mathbf{I}(1-2/n)+(1/n)\mathbf{1}\mathbf{1}^{\mathbf{T}}$, where $\mathbf{I},\mathbf{1}$ are respectively the identity matrix and the 2-column of ones. Thus, the HMM jumps to the alternative state after spending asymptotically exponentially distributed time with mean n in each state. Suppose for simplicity that we know the initial state (1).

The corresponding SCOT(1) structure is estimated following [10] during sufficiently small initial period of length $\alpha n, 0<\alpha<1$, which is also the starting interval of sequential detection of the first jump.

ii. Sequential detection of the change point.

The online detection of the first jump moment τ combines an initial online estimation and consequent offline update following section 3.1.i.

Numerous references deal with online change point detection: control charts, CUSUM, [1], etc. A recent survey is [16]. Here we *sketch a simple suboptimal approach of* [1] postponing optimization until future publications. Our SCOT log-likelihood process l_i converges weakly to the Brownian motion under broad conditions, see [9, 11]. Thus, applying the profound theories of [15] makes the development of more asymptotically sound online evaluation of $\hat{\tau}$ promising. Notice that our online change point detection serves for training alternative SCOT. After that it is updated by the retrospective more accurate estimate of section 3.1.i. Every step of this procedure is verified by SCOT homogeneity tests.

Denote by $P_\tau(E_\tau)$ distributions (expectations) corresponding to strings with change point τ and by $P_\infty(E_\infty)$ strings without change point.

A stopping time $\hat{\tau}_N$ based on N measurements is the first hitting time of certain set. Introduce non-negative part y^+ of $y\in R$ and delay time $(\hat{\tau}_N-\tau)^+/N$.

Consider

$$Y_N(k,r)=k^{-1}\sum_{r-N+1}^{r-N+k} l_i-(N-k)^{-1}\sum_{r-N+k+1}^{r} l_i,\quad k=1,\dots,N-1,\quad r=N,\dots,$$

r, N are running indexes of observations until a change point is detected.

Define for $0<\alpha<1$,

$$z_N(r)=\max|Y_N(k,r)|:\alpha N\le k\le(1-\alpha)N.$$

Their stopping time $\hat{\tau}$ is the hitting time of the region $z_N(r)\ge c$.

Under natural regularity conditions, the following bounds are proved showing consistency of $\hat{\tau}$

$$\limsup\left(E[(\hat{\tau}-\tau)|\hat{\tau}\ge\tau]\right)/\log E_\infty[\hat{\tau}-N]\le const;$$

$$E_\infty(\hat{\tau}-N)\ge A\exp(BN)(1+o(1)).$$

The last inequality shows that the Probability of false alarm far from τ is small.

The estimate $\hat{\tau}$ is used as an initial consistent approximation for τ which is updated offline in the next section.

iii. SCOT(± 1) training update and segmentation of the state constancy.

We train SCOT(-1) during small consequent interval of length αn belonging to the next period of constant HMM state. Next, we update the estimate offline according to the procedure described in section 3.1.i. and widen periods of SCOT(± 1) constancy omitting only small neighborhoods of current $\hat{\tau}$ estimates. These are applied for further updating SCOT(± 1). This alternating procedure is repeated in all constancy periods of HMM until convergence.

iv. Unknown HMM parameters estimation.

If it is only known that both time means spent in two states before jump are proportional to large parameter n, 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 jump moments statistics using obvious delay frequencies. Namely, denoting n_{ij} = number of times i is followed by $j, j = \pm 1$, under $P(X_1 = 1) = 1$, the log-likelihood is

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

yields

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

Thus, empirical time mean before estimated jump from i to $1 - i, i = \pm 1$, serves as an estimate of mean time spent in i, while transition probabilities are estimated via the last formula.

5. HMM WITH FINITE NUMBER OF STATES

Here we outline training SCOT and 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. It is followed by the SCOT training of the string after jump of length αn where homogeneity is verified by homogeneity test and by the subsequent 'straight line cross' offline change point update (section 3.1.i.) of the change point preliminary online estimate.

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

6. DISCUSSION, OPEN PROBLEMS AND ACKNOWLEDGMENTS

The training algorithm exposed in present paper relies heavily on repeated application of the SCOT training and homogeneity test developed by us earlier for a cluster of computers. This super-fast parallel evaluation makes the whole procedure viable.

Numerical implementation of the procedure on statistically simulated and real world examples will be exposed in our future publications.

Our next goal is also developing online SCOT training and homogeneity test parallel procedures for multi-channel online change point detection.

REFERENCES

1. Brodsky, B.E. and Darkhovsky, B.S., *Nonparametric methods in change-point problems*, Dodrecht: Kluwer, 1993.

2. Cappe, O., Moulines, E., and Rydyen, T., *Inference in Hidden Markov models*, New York: Springer, 2005.

3. Dumont, T., Context Tree estimation in Variable Length Hidden Markov Models, *IEEE Trans. Inform. Theory*, 2014.

4. Durbin R, Eddy S, Krogh A, Mitchison G. *Biological Sequence Analysis*, Cambridge University Press, 1998.

5. 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.

6. Hamiltons J. *Regime switching models*, The New Palgrave Dictionary of Economics. Second Edition, Eds. Steven N. Durlauf and Lawrence E. Blume, Palgrave Macmillan, 2008. doi:10.1057/9780230226203.1411.

7. 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.

8. M. B. Malyutov, T. Zhang, X. Li and Y. Li, Time series homogeneity tests via VLMC training, *Information Processes*, 2013, vol. 13, no. 4, pp. 401–414.

9. Malyutov M. B. and Zhang T. Limit theorems for additive functions of SCOT trajectories, *Information Processes*, 2015, vol. 15, no. 1, pp. 89-96.

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

11. Malyutov M. SCOT approximation and asymptotic inference I, *Information Processes*, 2017, vol. 17, no. 1, pp. 61–69.

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

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

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

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

16. 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.

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

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