Partially Markov models and unsupervised segmentation of semi-Markov chains hidden with long dependence noise

Jérôme Lapuyade-Lahorgue and Wojciech Pieczynski

GET/INT, CITI Department,
CNRS UMR 5157
9, rue Charles Fourier,
91000 Evry, France

Abstract. The hidden Markov chain (HMC) model is a couple of random sequences (X, Y), in which X is an unobservable Markov chain, and Y is its observable noisy version. Classically, the distribution $p(y|x)$ is simple enough to ensure the Markovianity of $p(x|y)$, that enables one to use different Bayesian restoration techniques. HMC model has recently been extended to "triplet Markov chain" (TMC) model, which is obtained by adding a third chain U and considering the Markovianity of the triplet $T = (X, U, Y)$. When U is not too complex, X can still be recovered from Y. In particular, a semi-Markov hidden chain is a particular TMC. Otherwise, the recent triplet partially Markov chain (TPMC) is a triplet $T = (X, U, Y)$ such that $p(x, u|y)$ is a Markov distribution, which still allows one to recover X from Y.
The aim of this paper is to introduce, using a particular TPMC, semi-Markov chains hidden with long dependence noise. The general iterative conditional estimation (ICE) method is then used to estimate the model parameters, and the interest of the new model in unsupervised data segmentation is validated through experiments.
Keywords: hidden Markov chains, triplet Markov chains, unsupervised segmentation, image segmentation, iterative conditional estimation.

1. Introduction

Let $X = (X_n)_{1 \le n \le N}$ and $Y = (Y_n)_{1 \le n \le N}$ two stochastic processes, where X is hidden and Y is observable. In the whole paper, each X_n takes its values in a finite set of classes $\Omega = \{\omega_1, \ldots, \omega_K\}$ and each Y_n takes its values in R. The problem of estimating X from Y, which occurs in numerous applications, can be solved with Bayesian methods once one has chosen some accurate distribution $p(x, y)$ for $Z = (X, Y)$. The hidden Markov chain (HMC) model is the simplest and most well known model. Its applications cover numerous fields, and can be seen in recent books or general papers [Koski, 2001], [Cappé *et al.*, 2005]. [Ephraim and Merhav, 2002]. However, it is insufficient in some situations and thus it has been extended to "hidden semi-Markov chains" models [Faisan et al., 2005], [Guédon, 2003], [Moore and Savic, 2004], [Yu and Kobayashi, 2003]. Otherwise, a long dependence noise does exist in some situations [Doukhan *et al.*, 2003], but can not be taken into account in the classical HMC. The aim of this paper is to propose a new model in which the hidden chain is a Markov one, and in which the noise is a long dependence one. On the one hand, we exploit the fact, already mentioned in [Pieczynski and Desbouvries, 2005] that an HSMC is a particular "triplet Markov chain" (TMC, [Pieczynski et al., 2002], [Ait-el-Fquih and Desbouvries, 2006]). On the other hand, we exploit the ideas proposed in [Pieczynski, 2004]. We also propose a parameter estimation method of "iterative conditional estimation" (ICE) kind [Fjortoft et al., 2003], and show that the new model can be of non negligible interest in unsupervised data segmentation.

2. Triplet Markov chains and hidden semi-Markov chains

Let $X = (X_n)_{1\le n\le N}$ and $Y = (Y_n)_{1\le n\le N}$ be two stochastic processes mentioned above. The problem of estimating $X = x$ from $Y = y$ can be solved once the marginal posterior distributions $p(x_n|y)$ are calculable. Let us consider an auxiliary process U taking its values in a finite state space $\Lambda = \{1,...,L\}$ and such that (X,U) is a Markov chain whose distribution, given by

$p(x,u) = p(x_1)p(u_1|x_1)\prod_{n=1}^{N-1} p(x_{n+1}|x_n,u_n)p(u_{n+1}|x_n,u_n,x_{n+1})$, verifies

$$p(x_{n+1}|x_n,u_n) = \delta_{x_n}(x_{n+1}) \text{ if } u_n > 1, \text{ and } p(x_{n+1}|x_n) \text{ if } u_n = 1;$$
$$p(u_{n+1}|x_{n+1},x_n,u_n) = \delta_{u_n-1}(u_{n+1}) \text{ if } u_n > 1, \text{ and } p(u_{n+1}|x_{n+1}) \text{ if } u_n = 1 \qquad (1)$$

The chain X is then called "semi-Markov chain". If we consider $p(y|x) = \prod_{n=1}^{N} p(y_n|x_n)$, the triplet (X,U,Y) is the classical "hidden semi-Markov chain" (HSMC). In this paper, we consider a more sophisticated noise distribution $p(y|x)$, which is a "long dependence" one. To introduce it, let us consider stationary Gaussian process $Y = (Y_1,...,Y_N)$. It will be called "long-dependence" if its covariance function $\gamma(k) = E(Y_nY_{n-k}) - E(Y_n)(Y_{n-k})$ is such that there exist $\alpha \in]0,1]$ and C for which $\gamma(k) \sim Ck^{-\alpha}$ when $|k| -> \infty$.

The new model $T=(X,U^1,U^2,Y)$ we propose is the following. (X,U^1) is a semi-Markov chain defined by (1), with $\Lambda^1=\{1,...,L_1\}$. The chain U^2 is such that each U_n^2 takes its values in $\Lambda^2=\{1,...,L_2\}$ and, at each $n=1$, …, N, the variable U_n^2 designates the number $k\le L_2$ of previous indices $n-1$, …, $n-k$ such that $x_n=x_{n-1}=...=x_{n-k}$, and $x_n\neq x_{n-k-1}$. Therefore we can say that U_n^2 is the exact past sojourn time in x_n, while U_n^1 is, according to (1), a minimal future time sojourn in x_n (in our model, $u_n^1=1$ does not imply that $p(x_{n+1}=x_n|x_n)=0$). We can note that when $x_{n+1}=x_n$, $u_{n+1}^1=u_n^1-1$ (if $u_{n+1}^1>1$) and $u_{n+1}^2=u_n^2+1$. Otherwise, as the noise is a long dependence one, the distribution $p(y|x)$ is not a Markov one. In the model we propose, the distribution of Y_{n+1} conditional on $X_{n+1}=x_{n+1}$, $U_{n+1}^2=k$, and $Y_1=y_1$, …, $Y_n=y_n$, depends on $X_{n+1}=x_{n+1}$ and $Y_{n-k}=y_{n-k}$, …, $Y_n=y_n$. Therefore, for each class x_n, the distribution $p(y_{n-u_n^2},y_{n-u_n^2+1},...,y_n|x_n)$ is Gaussian with the mean vector $M_{x_n}^{(u_n^2+1)}=(\underbrace{M_{x_n},....,M_{x_n}}_{u_n^2+1\text{ times}})$ and the variance-covariance matrix $\Gamma_{x_n}^{(u_n^2+1)}$ such that $(\Gamma_{x_n}^{(u_n^2+1)})_{i,j}=\sigma_{x_n}^2(|i-j|+1)^{-\alpha_{x_n}}$.

Finally, the distribution of the new model $T=(X,U^1,U^2,Y)$ we propose is defined by

$$p(x,u^1,u^2,y)=p(x_1)p(u_1^1|x_1)p(u_1^2|x_1)p(y_1|x_1)\times$$

$$\times\prod_{n=1}^{N-1}p(x_{n+1}|x_n,u_n^1)p(u_{n+1}^1|x_{n+1},u_n^1)p(u_{n+1}^2|x_n,x_{n+1},u_n^2)p(y_{n+1}|x_{n+1},u_{n+1}^2,y^n)$$

where:

$p(x_{n+1}|x_n,u_n^1)=\delta_{x_n}(x_{n+1})$ if $u_n^1>1$, and $p(x_{n+1}|x_n)$ if $u_n^1=1$;

$p(u_{n+1}^1|x_{n+1},u_n^1)=\delta_{u_n^1-1}(u_{n+1}^1)$ if $u_n^1>1$, and $p(u_{n+1}^1|x_{n+1})$ if $u_n^1=1$;

$p(u_{n+1}^2|x_n,x_{n+1},u_n^2)=\delta_{u_n{}^2+1}(u_{n+1}^2)$ if $x_n=x_{n+1}$, $u_n^2<L_2$, and $\delta_0(u_{n+1}^2)$ if $x_n\neq x_{n+1}$ or $u_n^2=L_2$;

$p(y_{n+1}|x_{n+1},u_{n+1}^2,y^n)=p(y_{n+1}|x_{n+1})$ if $u_{n+1}^2=0$, and $p(y_{n+1}|x_{n+1},y_{n-u_{n+1}^2+1},...,y_n)$ if $u_{n+1}^2\geq 1$.

Moreover, in the last relation the Gaussian vector $p(y_{n-u_{n+1}^2+1}, y_{n-u_{n+1}^2+2}, ..., y_{n+1} | x_{n+1})$ verifies long memory condition above.
It is then possible to show that for observed $Y = y$, the distribution $p(x, u^1, u^2 | y)$ is a Markov distribution and, as (X, U^1, U^2) is finite, the classical "Backward" and "Forward" calculations give $p(x_n, u_n^1, u_n^2 | y)$, which gives $p(x_n | y) = \sum_{u_n^1, u_n^2} p(x_n, u_n^1, u_n^2 | y)$ used in Bayesian MPM segmentation.

3. Parameter estimation with ICE

The "Iterative Conditional Estimation" (ICE) method we use in this paper is based on the following principle [Fjørtoft *et al.*, 2003]. Let $\theta = (\theta_1, ..., \theta_m)$ be the vector of all real parameters defining the distribution $p(t)$ of a TMC $T = (X, U, Y)$, and let $\hat{\theta}(t)$ be an estimator of θ defined from the complete data $t = (x, u, y)$. ICE is an iterative method consisting on:
(i) initialize θ^0;
(ii) compute $\theta_i^{q+1} = E[\hat{\theta}_i(X, U, Y) | Y = y, \theta^q)$ for the components θ_i for which this computation is workable;
(iii) for other components θ_i, simulate (x_1^q, u_1^q), …, (x_l^q, u_l^q) according to $p(x, u | y, \theta^q)$ and put $\theta_i^{q+1} = \frac{\hat{\theta}(x_1^q, u_1^q, y) + ... + \hat{\theta}(x_l^q, u_{l1}^q, y)}{j}$.
We see that ICE is applicable under very slight two hypotheses: existence of an estimator $\hat{\theta}(t)$ from the complete data, and the ability of simulating (X, U) according to $p(x, u | y)$. The first hypothesis is not really a constraint because if we are not able to estimate θ from complete data (x, u, y), there is no point in searching an estimator from incomplete ones y. The second hypothesis is always verified for any TMC $T = (X, U, Y)$; in fact, $p(x, u | y)$ is a Markov chain distribution.
In order to detail how ICE is working let us specify the different parameters and the possibility of their estimation from the complete data $T = (X, U^1, U^2, Y)$. First, the distribution $p(x_1, u_1^1)$ and the transitions of the Markov chain (X, U^1) are defined by $p(x_1, u_1^1, x_2, u_2^1)$ which we propose to estimate from $T = (X, U^1, U^2, Y)$ by the classical counting estimator (the function I is defined by $I(a = b) = 1$ if $a = b$, and 0 otherwise)

$$\hat{p}(x_1, u_1^1, x_2, u_2^1) = \frac{1}{N-1}\sum_{n=1}^{N-1} I(x_n = x_1, u_n^1 = u_1^1, x_{n+1} = x_2, u_{n+1}^1 = u_2^1) \quad (2)$$

Second, the "noise parameters" are the means M_{x_n}, the variances $\sigma^2{}_{x_n}$, and the long dependence parameters α_{x_n}, which we propose to estimate from $T = (X, U^1, U^2, Y)$ by

$$\hat{M}_x = \frac{\sum_{n=1}^{N} y_n I(x_n = x)}{I(x_n = x)}, \quad \hat{\sigma}^2{}_x = \frac{\sum_{n=1}^{N}(y_n - \hat{M}_x)^2 I(x_n = x)}{I(x_n = x)} \text{ and } \hat{\alpha}_x = -\frac{\log(\hat{\gamma}_x(1))}{\log(2)}$$

$$\text{where } \hat{\gamma}_x = \frac{\sum_{n=1}^{N-1}(y_n - m_x^1)(y_n - m_x^1) I(x_n = x, x_{n+1} = x)}{I(x_n = x, x_{n+1} = x)} \text{ with} \quad (3)$$

$$m_x^1 = \frac{\sum_{n=1}^{N-1} y_n I(x_n = x, x_{n+1} = x)}{I(x_n = x, x_{n+1} = x)} \text{ and } m_x^2 = \frac{\sum_{n=1}^{N-1} y_{n+1} I(x_n = x, x_{n+1} = x)}{I(x_n = x, x_{n+1} = x)}.$$

Concerning (2), then the computation of (ii) in ICE is possible and gives:

$$p^{(q+1)}(x_1, u_1^1, x_2, u_2^1) = \frac{1}{N-1}\sum_{n=1}^{N-1} p(x_n = x_1, u_n^1 = u_1^1, x_{n+1} = x_2, u_{n+1}^1 = u_2^1 \mid y, \theta_q),$$

where $p(x_n = x_1, u_n^1 = u_1^1, x_{n+1} = x_2, u_{n+1}^1 = u_2^1 \mid y, \theta_q)$ are computed using the forward-backward algorithm. Concerning the noise parameters, the conditional expectation is not computable and we have to use (iii). In experiments below the initialization is obtained from the segmentation by the classical k-means method, and we take $l = 1$.

4. Experiments

The new "hidden semi-Markov chains with long dependence noise" (HSMC-LDN) model generalizes, on the one hand, the classical "hidden semi-Markov chains" (HSMC) and, on the other hand, the "hidden Markov chains with long dependence noise" (HMC-LDN), which are a particular case of HSMC-LDN such that X is a Markov chain. The aim of this section is to test the interest of these two generalizations in unsupervised data segmentation framework.
To illustrate the results we will use images of size $N = 128 \times 128$. Such a bi-dimensional set of pixels is transformed into a mono-dimensional set using a Hilbert-Peano scan [Fjortoft *et al.*, 2003], which gives a mono-dimensional chain. Such a representation is quite pleasant because it allows one to

appreciate visually the degree of the noise, and also the quality difference between two segmentation results.

We present three series of results. In the first series, the data suit SHMC and the question is whether using the new more complex HSMC-LDN does not degrade the results. The second series is devoted to the converse problem: when data suit HSMC-LDN, how do SHMC and HMC-LDN work? Finally, in the third series we use data produced by non one of the three models.

Let (X, U^1, Y) be a classical SHMC, with $L_1 = 10$, means equal respectively to 1 and 2, and variances equal to 20. The distribution of $p(u_{n+1} | x_{n+1}, u_n = 1)$ is uniform on Λ^1, and $p(x_n, x_{n+1} | u_n = 1) = 0.4995$ for $x_n = x_{n+1}$, $p(x_n, x_{n+1} | u_n = 1) = 0.0005$ for $x_n \neq x_{n+1}$. The obtained realisation $Y = y$, presented in Fig.1, is then segmented by three methods. The first one is the MPM method based on true parameters; thus the result is the reference one. The second method is the MPM unsupervised method based on the classical HSMC and ICE, while the third method is the MPM unsupervised method based on the new HSMC-LDN model, with $L_2 = 50$, and the related ICE. The aims of this experiment are, on the one hand, to show the robustness of the HSMC-LDN model and on the other hand, to see how the new model manages the independent noise.

According to the results presented in Fig.1, we see that the new model gives comparable results. This is due to the good behaviour of the parameter estimation method; in fact, The estimates of means are 1.01, 2.04 for HSMC and 0.98, 1.97 for HSMC-LDN. The estimates of variances are 19.81, 20.71 for HSMC and 19.84, 20.46 for HSMC-LDN. Finally, the estimates of α are, for HSMC-LDN, 15.28 and 5.95.

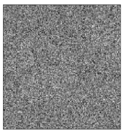
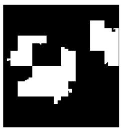
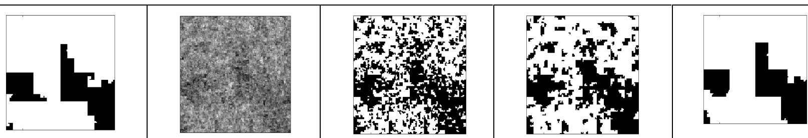
				
$X = x$	$Y = y$	MPM 3.74%	HSMC 4.55%	HSMC-LDN 4.57%

Fig. 1 Segmentation of the HSMC model according to three methods.

Let us now describe the second series of experiments. Here, we aim to segment a HSMC-LDN which is neither a particular SHMC nor a particular HMC-LDN. The the semi-Markov chain (X, U^1) is the same as above. For the noise, the means are respectively equal to 1 and 2, the variance is equal to 1 and $\alpha = 0.5$.

According to Fig. 2 we see that neither HSMC nor HMC-LDN can compete with HSMC-LDN when data suit the latter. The difference in error ratios is very large, which means that HSMC-LDN is a really significative extension of both HMC-LDN and HSMC.

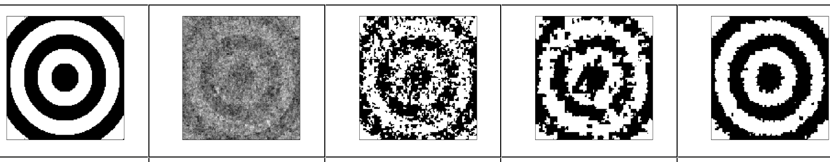


$X = x$	$Y = y$	HSMC 31.15%	HMC-LDN 21.53%	HSMC-LDN 3.16%

Fig. 2. Three unsupervised segmentations of data simulated according to HSMC-LDN

In this second example, the estimates of means are 0.97, 2.44 for HSMC, 1.08, 2.22 for HMC-LDN, and 0.98, 1.97 for HSMC-LDN. The estimates of variances are 0.59, 0.56 for HSMC, 0.83, 0.78 for HMC-LDN, and 0.96, 0.93 for HSMC-LDN. Finally, the estimates of α are 0.69, 0.72 for HMC-LDN and 0.62, 0.61 for HSMC-LDN.

Finally, we consider a hand-written image $X = x$ presented in Fig. 3. The means of the noise are respectively equal to 1 and 2, whereas the common variance is equal to 1. As above, we are segmenting $Y = y$ by using the three methods SHMC, HMC-LM and SHMC-LM. As above, we consider that $L_1 = 10$ for the semi-markovianity and $L_2 = 50$.



$X = x$	$Y = y$	HSMC 24.70%	HMC-LDN 18.27 %	HSMC-LDN 6.31 %

Fig. 3. Three unsupervised segmentations of hand written $X = x$ noisy with a long dependence noise.

The estimates of means are 0.75, 2.26 for HSMC, 0.91, 2.04 for HMC-LDN, and 0.99, 1.99 for HSMC-LDN. The estimates of variances are 0.69, 0.66 for HSMC, 0.92, 0.93 for HMC-LDN, and 1.01, 1.02 for HSMC-LDN. Finally, the estimates of α are 1.07, 1.07 for HMC-LDN and 0.93, 0.92 for HSMC-LDN.

5. Conclusion

As a general conclusion we can say, accordingly to different experiments results, that the new semi-Markov chain hidden with long dependence noise model proposed in this paper turns out to be of interest, when unsupervised segmentation is concerned, with respect to classical simpler models.

References

[Cappé *et al.*, 2005] O. Cappé, E. Moulines, T. Ryden, *Inference in hidden Markov models,* Springer, Series in Statistics, 2005.

[Doukhan *et al.*, 2003],] P. Doukhan, G. Oppenheim, and M. S. Taqqu, editors, Theory and Applications of Long-Range Dependence, Birhhäuser, 2003.

[Ephraim and Merhav, 2002] Y. Ephraim and N. Merhav, Hidden Markov processes, *IEEE Trans. on Information Theory,* Vol. 48, No. 6, pp. 1518-1569, 2002.

[Faisan *et al.*, 2005] S. Faisan, L. Thoraval, J.-P. Armspach, M.-N. Metz-Lutz , F. Heitz, Unsupervised learning and mapping of active brain functional MRI signals based on hidden semi-Markov event sequence models, *IEEE Trans. On Medical Imaging,* Vol. 24, No. 2, pp. 263-276, 2005.

[Fjørtoft *et al.*, 2003] R. Fjørtoft, Y. Delignon, W. Pieczynski, M. Sigelle, and F. Tupin, Unsupervised segmentation of radar images using hidden Markov chains and hidden Markov random fields, *IEEE Trans. on Geoscience and Remote Sensing*, Vol. 41, No. 3, pp. 675-686, 2003.

[Guédon, 2003] Y. Guédon, Estimating Hidden Semi-Markov Chains from discrete Sequences, Journal of Computational and Graphical Statistics, Vol. 12, No. 3, pp. 604-639, Sept. 2003.

[Koski, 2001] T. Koski, Hidden Markov models for bioinformatics, Kluwer Academic Publishers, 2001.

[Moore and Savic, 2004] M. D. Moore and M. I. Savic, Speech reconstruction using a generalized HSMM (GHSMM), Digital Signal Processing, Vol. 14, No. 1, pp. 37-53, 2004.

[Pieczynski *et al.*, 2002] W. Pieczynski, C. Hulard, and T. Veit, Triplet Markov Chains in hidden signal restoration, *SPIE's International Symposium on Remote Sensing*, September 22-27, Crete, Greece, 2002.

[Pieczynski, 2004] W. Pieczynski, Triplet Partially Markov Chains and Tree, 2nd International Symposium on Image/Video Communications over fixed and mobile networks (ISIVC'04), Brest, France, 7-9 July, 2004.

[Pieczynski and Desbouvries, 2005] W. Pieczynski and F. Desbouvries, On triplet Markov chains, International Symposium on Applied Stochastic Models and Data Analysis, (ASMDA 2005), Brest, France, May 2005.

[Yu and Kobayashi, 2003] S.-Z. Yu and H. Kobayashi, A hidden semi-Markov model with missing data and multiple observation sequences for mobility tracking, *Signal Processing*, Vol. 83, No. 2, pp. 235-250, 2003.