Performance comparison across hidden, pairwise and triplet Markov models' estimators

I. GORYNIN, L. CRELIER, H. GANGLOFF, E. MONFRINI and W. PIECZYNSKI
SAMOVAR, Telecom SudParis, CNRS
Université Paris-Saclay
9, rue Charles Fourrier, 91000 Evry
FRANCE
ivan.gorynin@telecom-sudparis.eu http://citi.telecom-sudparis.eu/fr/

Abstract: In a hidden Markov model (HMM), one observes a sequence of emissions (**Y**) but lacks a Markovian sequence of states (**X**) the model went through to generate these emissions. The hidden Markov chain allows recovering the sequence of states from the observed data. The classic HMM formulation is too restrictive but extends to the *pairwise* Markov models (PMMs) where we assume that the pair (**X**, **Y**) is Markovian. Since (**X**) is not necessarily Markovian in such a model, a PMM is generally not a *hidden* Markov model. However, since (**X**) is conditionally Markovian in the PMM, an HMM-like fast processing is available. Similarly, the *triplet* Markov models (TMMs) extend the PMMs by introducing an additional unobserved discrete-valued process (**U**). The triplet (**X**, **U**, **Y**) is Markovian in a TMM. Such a model is more inclusive than the PMM and offers the same possibilities of fast processing. The aim of the paper is to present numerical studies which evaluate how these models may behave compared to the classic HMM. In other words, we compare different models in terms of the Bayesian Maximum Posterior Mode (MPM) error rate. We show that the misclassification percentage decreases by a half when using these advanced models.

Key-Words :— Bayesian classification, Hidden Markov models, Maximum Posterior Mode, PMM, TMM.

1 Introduction

The *hidden* Markov model HMM [1,2] is an important tool in the modern modelling [3-7]. Let us consider a hidden random sequence $\mathbf{X} = (X_1, ..., X_N)$ and an observed one $\mathbf{Y} = (Y_1, ..., Y_N)$, which we describe by using the probability density function $p(\mathbf{x}, \mathbf{y})$ of $(\mathbf{X}, \mathbf{Y})$. Each X_n takes its values in $\Omega = \{1, ..., K\}$ and each Y_n takes its values in R . The pair $(\mathbf{X}, \mathbf{Y})$ is a classic HMM if and only if

(i) $\mathbf{X}$ is a Markov chain;

(ii) $p(y_1 ... y_N | x_1 ... x_N) = p(y_1 | x_1) ... p(y_N | x_N)$.

The *pairwise* Markov models (PMMs) extend these models by assuming that $(\mathbf{X}, \mathbf{Y})$ is Markovian [8,9]. Since in the PMMs the hidden process $\mathbf{X}$ is not necessarily Markov, they are strictly more general than HMMs. We note that in the PMM framework, a HMM-like processing is available.

Next, the *triplet* Markov models (TMMs) extend the PMMs by using a discrete-valued latent process $\mathbf{U} = (U_1, ..., U_N)$ where each U_n belongs to a finite set $\Lambda = \{1, ..., M\}$. In such a model, $\mathbf{T} = (\mathbf{X}, \mathbf{U}, \mathbf{Y})$ is Markovian. Similarly to both PMMs and HMMs, any HMM-like processing is available in TMMs, even if the processes $\mathbf{X}$, $\mathbf{U}$, $\mathbf{Y}$, $(\mathbf{X}, \mathbf{U})$, $(\mathbf{X}, \mathbf{Y})$, and $(\mathbf{U}, \mathbf{Y})$ are not necessarily Markovian.

The problem we focus on in this paper consists in exploring if using TMMS instead of PMMs is meaningful for practical applications. The next section is devoted to the PMMs, while the third one presents the TMMs. Conclusions and perspectives are in section 4.

2 Pairwise Markov models

2.1 Bayesian segmentation using pairwise Markov models

The pair $(\mathbf{X}, \mathbf{Y})$ is a pairwise Markov model (PMM) if its distribution $p(\mathbf{x}, \mathbf{y})$ is

$$p(\mathbf{x}, \mathbf{y}) = \qquad (2.1)$$
$$p(x_1, y_1) p(x_2, y_2 | x_1, y_2) ... p(x_N, y_N | x_{N-1}, y_{N-1}) ,$$

which means that $(\mathbf{X}, \mathbf{Y})$ is Markovian. Since the classic HMM distribution is

$$p(\mathbf{x}, \mathbf{y}) = \qquad (2.2)$$
$$p(x_1) p(y_1 | x_1) p(x_2 | x_1) p(y_2 | x_2) ... p(x_N | x_{N-1}) p(y_N | x_N) ,$$

and the transitions $p(x_n, y_n | x_{n-1}, y_{n-1})$ in (2.1) are

$$p(x_n, y_n | x_{n-1}, y_{n-1}) = \\ p(x_n | x_{n-1}, y_{n-1}) p(y_n | x_n, x_{n-1}, y_{n-1}) \,, \tag{2.3}$$

a PMM is an HMM if and only if

$$p(x_n | x_{n-1}, y_{n-1}) = p(x_n | x_{n-1}) \text{ and} \tag{2.4}$$

$$p(y_n | x_n, x_{n-1}, y_{n-1}) = p(y_n | x_n) \,. \tag{2.5}$$

We use the maximum posterior mode (MPM) estimator [10], which estimates the state vector by $\hat{\mathbf{x}} = (\hat{x}_1, ..., \hat{x}_N)$ such that for each $n = 1, \ldots, N$:

$$\hat{x}_n = \arg\max_{k \in \Omega} p(x_n = k | \mathbf{y}). \tag{2.6}$$

Thus, we compute $p(x_n | \mathbf{y})$ for each $n = 1, \ldots, N$. Let us briefly recall the PMM-related forward-backward algorithm. The forward and backward probabilities $\alpha_n(x_n) = p(x_n, y_1, ..., y_n)$, $\beta_n(x_n) = p(y_{n+1}, ..., y_N | x_n, y_n)$ arise from the following recursions:

$$\alpha_1(x_1) = p(x_1, y_1) \,, \\ \alpha_{n+1}(x_{n+1}) = \sum_{x_n \in \Omega} p(x_{n+1}, y_{n+1} | x_n, y_n) \alpha_n(x_n) \,; \tag{2.7}$$

$$\beta_N(x_N) = 1 \,, \\ \beta_n(x_n) = \sum_{x_{n+1} \in \Omega} p(x_{n+1}, y_{n+1} | x_n, y_n) \beta_{n+1}(x_{n+1}) \,. \tag{2.8}$$

and then we have

$$p(x_n | \mathbf{y}) = \frac{\alpha_n(x_n) \beta_n(x_n)}{\sum_{x_n \in \Omega} \alpha_n(x_n) \beta_n(x_n)} \,. \tag{2.9}$$

Thus, the complexity of this algorithm is linear in n. The well known HMM version of this algorithm appears when we meet the conditions (2.4) and (2.5). Of course, we have $p(x_{n+1}, y_{n+1} | x_n, y_n) = p(x_{n+1} | x_n) p(y_{n+1} | x_{n+1})$ if $(\mathbf{X}, \mathbf{Y})$ is a HMM, what links this version of the algorithm to the conventional relations on forward-backward probabilities.

2.2 Stationary invertible PMMs

In what follows, we consider the stationary PMM for which $p(x_n, y_n, x_{n+1}, y_{n+1})$ does not depend on n. Thus, the whole distribution derives from $p(x_1, y_1, x_2, y_2)$. In addition, we assume that the transition kernel is invertible, i.e. $p(x_{n+1}, y_{n+1} | x_n, y_n) = p(x_n, y_n | x_{n+1}, y_{n+1})$.

Let us consider the following PMM sub-models:

(i) The genuine PMM in which the noise may be correlated and in which $\mathbf{X}$ may not be Markovian. We call it PMM with correlated noise (PMM-CN). The transitions $p(x_2, y_2 | x_1, y_1)$ are of the general form

$$p(x_2, y_2 | x_1, y_1) = p(x_2, | x_1, y_1) p(y_2 | x_1, y_1, x_2) \tag{2.10}$$

and the distribution $p(x_1, y_1, x_2, y_2)$ is also of the general form, since we have

$$p(x_1, y_1, x_2, y_2) = p(x_1, x_2) p(y_1, y_2 | x_1, x_2) \,; \tag{2.11}$$

(ii) PMM with independent noise (PMM-IN), where $\mathbf{X}$ might not be Markovian and where the observation noise is independent from $\mathbf{X}$. We have

$$p(x_2, y_2 | x_1, y_1) = p(x_2 | x_1, y_1) p(y_2 | x_1, x_2) \tag{2.12}$$

and

$$p(x_1, y_1, x_2, y_2) = p(x_1, x_2) p(y_1 | x_1, x_2) p(y_2 | x_1, x_2) \,. \tag{2.13}$$

We can show that Y_n and $(X_1, X_2, ..., X_{n-2})$ are independent conditional on X_{n-1} in a PMM-IN *cf.* Fig. 2, and the same holds for $(X_{n+2}, X_{n+3}, ..., X_N)$ at each $1 < n < N-1$. That is why the distribution of Y_n conditional on $\mathbf{X}$ is the same as the distribution of Y_n conditional on (X_{n-1}, X_n, X_{n+1}).

(iii) HMMs with correlated noise (HMM-CN). The related transition kernel is

$$p(x_2, y_2 | x_1, y_1) = p(x_2, | x_1) p(y_2 | x_1, x_2, y_1) \tag{2.14}$$

and $p(y_2 | x_1, x_2) = p(y_2 | x_2)$, which is not guaranteed to hold in the case of PMM-IN (2.11) (*see* Remark 1). We have

$$p(x_1, y_1, x_2, y_2) = p(x_1, x_2) p(y_1 | x_1, x_2) p(y_2 | x_1, x_2). \quad (2.15)$$

(iv) HMMs with independent noise, denoted with HMM-IN, which are the classic HMMs. The related transition kernel is

$$p(x_2, y_2 | x_1, y_1) = p(x_2 | x_1) p(y_2 | x_2) \quad (2.16)$$

and $p(x_1, y_1, x_2, y_2)$ verifies

$$p(x_1, y_1, x_2, y_2) = p(x_1, x_2) p(y_1 | x_1) p(y_2 | x_2) . \quad (2.17)$$

We supply the graphical representations of these sub-models in Fig. 1-4.

Remark 1

We show that $\mathbf{X}$ is Markovian in HMM-CN and HMM-IN. We also state that $\mathbf{X}$ is not necessarily Markovian in PMM-CN and PMM-IN by making use of a property that can be find in [9]. For a reversible and stationary pair $(\mathbf{X}, \mathbf{Y})$, the property is that $\mathbf{X}$ is Markovian if and only if the equation $p(y_2 | x_1, x_2) = p(y_2 | x_2)$ holds.

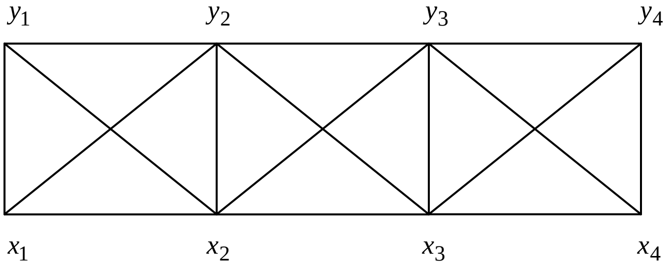

Fig. 1. The dependence graph of the PMM-CN.

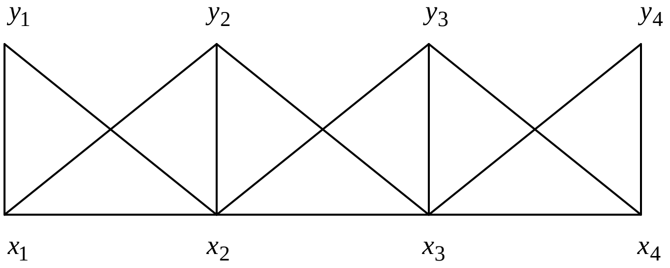

Fig. 2. The dependence graph of the PMM-IN.

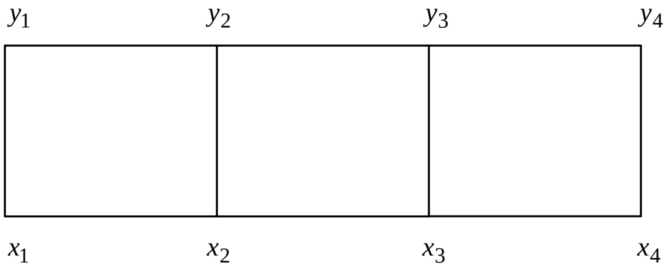

Fig. 3. The dependence graph of the HMM-CN.

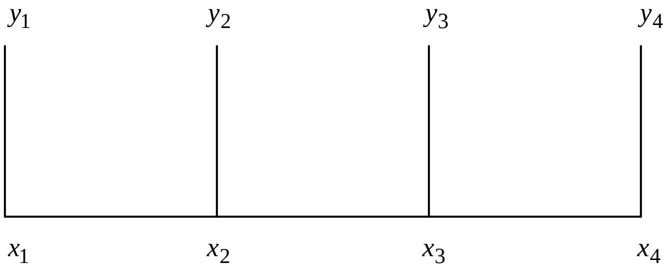

Fig. 4. The dependence graph of the HMM-IN.

3 Triplet Markov models

Let us consider hidden $\mathbf{X} = (X_1, ..., X_N)$ and observed $\mathbf{Y} = (Y_1, ..., Y_N)$ sequences as we did previously. The *triplet* Markov model (TMM) requires another discrete-valued process $\mathbf{U} = (U_1, ..., U_N)$, where each U_n belongs to a finite set $\Lambda = \{1, ..., M\}$. We suppose that the triplet $\mathbf{T} = (\mathbf{X}, \mathbf{U}, \mathbf{Y})$ is Markovian. The TMM is at a next level of evolution of HMM towards more inclusive models *cf.* Remark 2.

Similarly to HMM and PMM, the TMM allows recovering $\mathbf{X} = (X_1, ..., X_N)$ from $\mathbf{Y} = (Y_1, ..., Y_N)$ in a reasonable time. We do it thanks to the substitution $\mathbf{V} = (\mathbf{X}, \mathbf{U})$ which allows recovering $\mathbf{V}$ from $\mathbf{Y}$ in the $(\mathbf{V}, \mathbf{Y})$-PMM framework and then we extract $\mathbf{X}$ from $\mathbf{V}$. The TMMs have been successfully used to solve different problems in several domains [11]. Besides, $\mathbf{U}$ is an add-on latent variable to extend the PMM and may have a case-related interpretation. We believe that the TMMs have an incredible potential of modelization, where $\mathbf{U}$ would be multivariate like $(\mathbf{U}^1, ..., \mathbf{U}^s)$ so each sequence $\mathbf{U}^i$ models a separate property. For example, the non-stationary hidden semi-Markov chains can be seen as a TMM $\mathbf{T} = (\mathbf{X}, \mathbf{U}^1, \mathbf{U}^2, \mathbf{Y})$ in which $\mathbf{U}^1$ models the semi-Markovianity and $\mathbf{U}^2$ is for the non-stationarity [11].

Let us now consider a stationary invertible TMM with the distribution defined by

$$\begin{aligned} p(x_1,u_1,y_1,x_2,u_2,y_2) = \\ p(u_1,u_2)p(y_1|u_1)p(x_1|u_1)p(x_2|u_2)p(y_2|u_2). \end{aligned} \quad (3.1)$$

Then the transitions are

$$p(x_2,u_2,y_2|x_1,u_1,y_1) = p(u_2|u_1)p(x_2|u_2)p(y_2|u_2). \quad (3.2)$$

We call this model "simplified TMM" (STMM) and we observe that a STMM (*cf.* Fig. 5) is not a PMM; for example, we have $p(u_2|x_1,y_1,y_2,x_2) \neq p(u_2|y_2,x_2)$. We announced earlier that the objective of the paper was to provide a numerical comparison among different PMMs; however, we also provide a comparison between STMM and the classic HMM-IN.

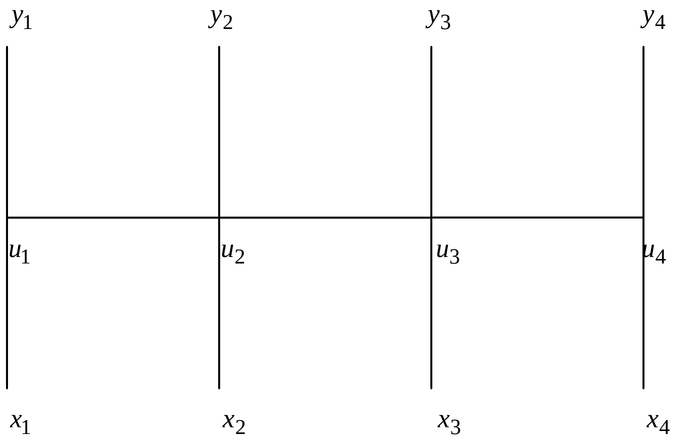

Fig. 5. The dependence graph of the STMM.

Remark 2

The TMMs are strictly more general than PMMs. In fact, one can use the result from [9] for the case of the stationary reversible systems. Let $\mathbf{T}=(\mathbf{X},\mathbf{U},\mathbf{Y})$ be a TMM; we set $\mathbf{Z}=(\mathbf{X},\mathbf{Y})$ to make appear $\mathbf{T}=(\mathbf{U},\mathbf{Z})$ as a PMM. Thus, $\mathbf{Z}=(\mathbf{X},\mathbf{Y})$ is Markovian if and only if $p(u_2|x_1,y_1,y_2,x_2) = p(u_2|y_2,x_2)$. Thus, we obtain a reversible stationary TMM $\mathbf{T}=(\mathbf{X},\mathbf{U},\mathbf{Y})$ with non Markovian $(\mathbf{X},\mathbf{Y})$ by choosing a distribution such that $p(u_2|x_1,y_1,y_2,x_2) \neq p(u_2|y_2,x_2)$.

4 Experiments

4.1 Pairwise Markov models

We present different experiments to compare PMMs-CN, PMMS-IN, HMMs-CN and HMMs-IN from Section II. We decide to set $\Omega = \{\omega_1, \omega_2\}$ for the sake of simplicity.

We begin with sampling data points of the most inclusive model (PMM-CN), for which we choose

$$p(x_1,x_2) = 1_{[x_1=x_2]}(0.5-\varepsilon) + 1_{[x_1\neq x_2]}\varepsilon;$$

$$p(y_1,y_2|x_1,x_2) = \mathrm{N}\left(\begin{bmatrix} \mu_1^{x_1,x_2} \\ \mu_2^{x_1,x_2} \end{bmatrix}, \begin{bmatrix} \left(\sigma_1^{x_1,x_2}\right)^2 & \rho \\ \rho & \left(\sigma_2^{x_1,x_2}\right)^2 \end{bmatrix}\right).$$

The coefficients ε (the probability of regime-switching) and ρ (the conditional correlation*)* depend on the experimental setting. The values of the remaining parameters are per each pair (x_1,x_1) are in the table below.

(x_1,x_1)	$\mu_1^{x_1,x_2}$	$\mu_2^{x_1,x_2}$	$\sigma_1^{x_1,x_2}$	$\sigma_2^{x_1,x_2}$
(ω_1,ω_1)	-5	-5	14	14
(ω_1,ω_2)	-3	3	7	9
(ω_2,ω_1)	3	-3	9	7
(ω_2,ω_2)	5	5	20	20

Next, we compute **X** from **Y** by using the PMM backward-forward algorithm for the "projections" of the known parameters into each sub-model. Finally, we compute a relative error rate, referring to the PMMs-CN (in percents):

$$\tau = \frac{\mathrm{err}_{\mathrm{model}} - \mathrm{err}_{\text{PMM-CN}}}{\mathrm{err}_{\text{PMM-CN}}}. \quad (4.1)$$

For example, if the relative error rate reaches 100%, then it means that the reference model decreases the misclassification percentage by a half when compared to the proposal one.

The values in Table 1 are the relative error rates (in percents) of the three models compared to PMMs-CN for various probabilities of regime-switching changes. Each value is averaged over the values of the correlation coefficient.

The values in Table 2 are the relative error rates of the three models compared to PMMs-CN for various correlation coefficients. Each value is averaged over the values of the probabilities of regime-switching.

We simulate 40 random chains for each pair of (ε, ρ) and 1000 elements per each chain.

ε	Model compared to PMM-CN		
	HMM-IN	HMM-CN	PMM-IN
0.025	40.7	4.9	35.7
0.125	34.9	11.0	18.5
0.275	55.5	13.6	21.1
0.425	57.2	13.3	48.8
Average	46.8	11.6	27.0

Table 1. Relative error rates (4.1) of the three models for varying probability of regime-switching.

ρ	Model compared to PMM-CN		
	HMM-IN	HMM-CN	PMM-IN
0.05	19.3	10.2	0.3
0.25	22.3	10.1	3.6
0.75	56.9	12.4	37.9
0.95	142.0	17.6	112.7
Average	60,1	12,6	38,6

Table 2. Relative error rates (4.1) of the three models for varying noise correlation coefficient.

These tables can be seen as “cuts” of the 3-D surface plot from Fig. 6.

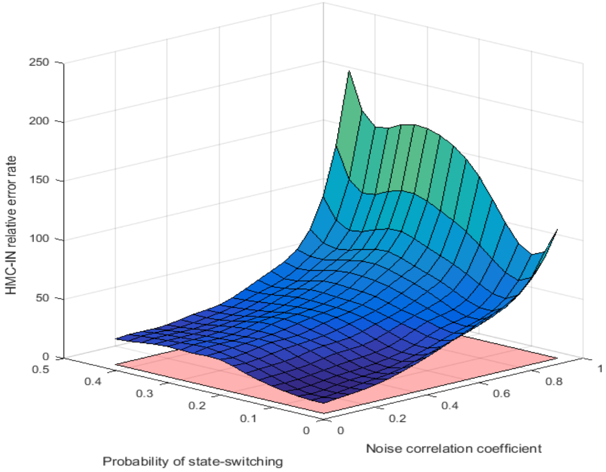

Fig. 6. Surface plot of the relative error rate. The height is a function which assigns to each pair *(ε,ρ)* the corresponding relative error (4.1).

4.2 STMM compared to the HMM-IN

In this sub-section we investigate if the STMM from Section 3 is competitive with the HMM-IN.

We decide to set $\Omega = \{\omega_1, \omega_2\}$ for $\mathbf{X}$, $\Lambda = \Omega$ for $\mathbf{U}$. We choose the state space distribution:

$$p(u_1, u_2) = 1_{[u_1 = u_2]} 0.49 + 1_{[u_1 \neq u_2]} 0.01;$$
$$p(x_1 | u_1) = 1_{[x_1 = u_2]} 0.99 + 1_{[x_1 \neq u_1]} 0.01;$$

Regarding the observation space, we have

$$p(y_1 | u_1 = \omega_1') = \mathrm{N}(y_1, 10, \sigma);$$
$$p(y_1 | u_1 = \omega_2') = \mathrm{N}(y_1, 20, \sigma).$$

Then we compute the error rates relative to the HHM-IN- and STMM-based MPM state estimations for various values of σ. We present our results in Fig. 7. They appear promising enough to be worth researching.

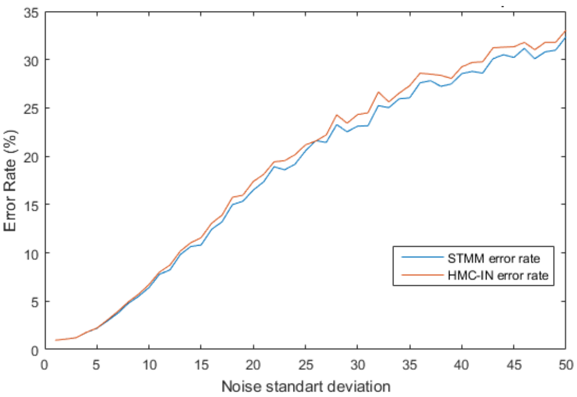

Fig. 7. Comparison between the performances of STMM and HMM-IN for various noise levels.

5 Conclusion

The primary objective of the paper is to compare efficiencies of Bayesian classifiers based on four different models: the classic HMM, the general PMM, and two intermediary models. Different results of experiments, some of which are presented in the paper, show that the PMMs potentially outperform the HMMs in “real-world” applications. Indeed, the PMM allows reducing the misclassification ratio by 10%-30% and even more. Such a gap is particularly visible if the observation noise is heavily correlated and if the hidden chain is too far from being Markovian. We also studied an example of a simplified triplet Markov chain, as simple as a HMM but very different from the latter.

The further work will include exploring more advanced TMMs, PMMs [12, 13] and their inter-comparisons with the on a similar methodology basis.

References:

[1] O. Cappé, E. Moulines, T. Rydén, Inference in Hidden Markov Models, Springer Verlag, 2005.

[2] Y. Ephraim and N. Merhav, Hidden Markov processes, IEEE Transactions on information theory, Vol. 48, No. 6, 2002, pp. 1518-1569.

[3] R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257-286.

[4] R. Bhar and S. Hamori, *Hidden Markov Models: Applications to Financial Economics*, Springer Science & Business Media, 2006

[5] R. Mamon and R. Elliott, *Hidden Markov Models in Finance*, Springer, 2007.

[6] T. Koski, *Hidden Markov Models for Bioinformatics*, *Computational Biology*, Springer Science & Business Media, 2001.

[7] M. Vidyasagar, *Hidden Markov Processes: Theory and Applications to Biology, Princeton Series in Applied Mathematics*, Princeton University Press, 2014.

[8] S. Derrode and W. Pieczynski, Unsupervised classification using hidden Markov chain with unknown noise copulas and margins, Signal Processing, vol. 128, 2016, pp. 8-17.

[9] S. Derrode and W. Pieczynski, Signal and image segmentation using Pairwise Markov Chains, IEEE Trans. on Signal Processing, Vol. 52, No. 9, 2004, pp. 2477-2489.

[10] L. Baum, T. Petrie, G. Soules and N. Weiss, A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, The annals of mathematical statistics, vol. 41, No. 1, 1970, pp. 164-171.

[11] P. Lanchantin, J. Lapuyade-Lahorgue and W. Pieczynski, Unsupervised segmentation of randomly switching data hidden with non-Gaussian correlated noise, Signal Processing, Vol. 91, No. 2, 2011, pp. 163-175.

[12] S. Carincotte, S. Derrode and S. Bourennane, Unsupervised change detection on SAR images using fuzzy hidden Markov chains, IEEE Transactions on Geoscience and Remote Sensing, Vol. 44, No. 2, 2006, pp. 432-441.

[13] F. Salzenstein, C. Collet, S. Cam and M. Hatt, Non-stationary fuzzy Markov chain, Pattern Recognition Letters, vol. 28, no 16, 2007, pp. 2201-2208.