Twenty six International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering, MaxEnt 2006, Paris, France, July 8-13, 2006

Unsupervised Segmentation of Hidden Semi-Markov Non Stationary Chains

Jérôme Lapuyade-Lahorgue and Wojciech Pieczynski

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

Abstract. In the classical hidden Markov chain (HMC) model we have a hidden chain X, which is a Markov one and an observed chain Y. HMC are widely used; however, in some situations they have to be replaced by the more general "hidden semi-Markov chains" (HSMC), which are particular "triplet Markov chains" (TMC) $T = (X, U, Y)$, where the auxiliary chain U models the semi-Markovianity of X. Otherwise, non stationary classical HMC can also be modeled by a triplet Markov stationary chain with, as a consequence, the possibility of parameters' estimation. The aim of this paper is to use simultaneously both properties. We consider a non stationary HSMC and model it as a TMC $T = (X, U^1, U^2, Y)$, where U^1 models the semi-Markovianity and U^2 models the non stationarity. The TMC T being itself stationary, all parameters can be estimated by the general "Iterative Conditional Estimation " (ICE) method, which leads to unsupervised segmentation. We present some experiments showing the interest of the new model and related processing in image segmentation area.

Key Words: Non-stationary hidden semi-Markov chain, unsupervised segmentation, iterative conditional estimation, triplet Markov chains.
PACS: 02.50.Ga

Notations: In this article, all the processes and random variables will be defined on the abstract probability space $(E, \xi, \Pr)$. The processes will be written in upper case letters and their realizations in lower case letters. The marginals will be indexed by the corresponding indexes. Except ambiguities $p(t|s)$ will be noted $\Pr(T = t|S = s)$ with the corresponding letters. If T is continuous, $p(t|s)$ will be a probability density function (pdf).

INTRODUCTION

In the classical hidden Markov chain (HMC) model there is a hidden random chain X, which is a Markov one, and an observed random chain Y. HMC are efficient and widely used in numerous problems; however, in some situations they have to be replaced by the more general "hidden semi-Markov chains" (HSMC) [3, 5, 7, 10, 11]. Otherwise, it has been recently showed that HSMC are particular "triplet Markov chains" (TMC [8]) $T = (X, U, Y)$, where an auxiliary chain U models the fact that X is semi-Markov [9]. Furthermore, it has been also showed that a non stationary classical hidden Markov chain can also be seen as a triplet Markov stationary chain with, as a consequence, the possibility of parameters estimation [6]. The aim of this paper is to use simultaneously both properties. We firstly consider a TMC $T^1 = (X, U^1, Y)$, which is equivalent to a hidden semi-Markov chain. Then we consider that T^1 is not stationary, which is modeled by a second auxiliary random chain U^2. Finally, we consider $T = (X, U^1, U^2, Y)$ as a TMC $T = (X, U, Y)$ with the auxiliary process $U = (U^1, U^2)$. Therefore we have a stationary TMC $T = (X, U, Y)$ which models a non stationary HSMC (NSHSMC). We propose to use such a TMC in unsupervised hidden discrete signal segmentation. The parameters estimation is performed by an original variant of the general "Iterative Conditional Estimation" (ICE) method [1, 2, 4], and the Bayesian segmentation is performed by the classical Maximum Posterior Mode (MPM) method. The interest of the new modeling and related processing is validated by some experiments.

MODELING HIDDEN NON STATIONARY SEMI-MARKOV CHAINS WITH TRIPLET MARKOV CHAINS

Let us consider $Z = (X, Y)$, with $X = (X_1, ..., X_n)$ and $Y = (Y_1, ..., Y_n)$ two random chains, where each X_i takes its values in a finite set of classes $\Omega = \{\omega_1, ..., \omega_K\}$, and each Y_i takes its values in R. Classically, $Z = (X, Y)$ is a hidden semi-Markov chain when X is a semi-Markov chain and when the distribution of Y conditional on X is given by $p(y|x) = p(y_1|x_1)...p(y_n|x_n)$. Otherwise, a possible way to define semi-Markov distribution of X is to say that this is a marginal distribution of a particular Markov chain. More precisely, one considers a random chain $U^1 = (U_1^1, ..., U_n^1)$, where each U_i^1 takes its values in the set of positive integers $N^* = \{1, 2, ...\}$, such that the couple (X, U) is a Markov chain defined by following $p(x_1, u_1^1)$ and transitions $p(x_{i+1}, u_{i+1}^1 | x_i, u_i^1)$. For each $i = 1, \ldots, n$ and $x_i \in \Omega$, one considers a probability distribution $p(.|x_i)$ on N^*, such that for $j \in N^*$, $p(j|x_i)$ is the probability that $(X_{i+1}, ..., X_{i+j}) = (x_i, ..., x_i)$, and $x_{i+1} \neq x_i$. This models the fact that the distribution of the "sojourn time" of the chain X in a given state can be of any form, while it is necessarily of geometrical form in Markov chains. More precisely, a semi-Markov distribution of X is the marginal distribution of a Markov chain (X, U^1) whose distribution is given by $p(x_1, u_1^1)$ and the following transitions $p(x_{i+1}, u_{i+1}^1 | x_i, u_i^1) = p(x_{i+1} | x_i, u_i^1) p(u_{i+1}^1 | x_{i+1}, x_i, u_i^1)$:

$$p(x_{i+1} | x_i, u_i^1) = \delta_{x_i}(x_{i+1}) \text{ if } u_i^1 > 1, \text{ and}$$

$$p(x_{i+1} | x_i) \text{ if } u_i^1 = 1 \text{ (with } p(x_{i+1} = x_i | x_i) = 0\text{);} \qquad (1)$$

$$p(u_{i+1}^1 | x_{i+1}, x_i, u_i^1) = \delta_{u_i - 1}(u_{i+1}) \text{ if } u_i^1 > 1, \text{ and } p(u_{i+1}^1 | x_{i+1}) \text{ if } u_i^1 = 1 \text{ ;} \qquad (2)$$

where δ_x is the Dirac measure on x. Let us notice that the variable $U_i^1 = u_i^1$ designates the remaining sojourn time in the state x_i.

Returning to the observed Y, the distribution of a hidden semi-Markov chain $Z = (X, Y)$ is the marginal distribution of a particular triplet Markov chain $T^1 = (X, U^1, Y)$. Let us put $V = (X, U^1)$. As V is a Markov chain, $T^1 = (V, Y)$ is a hidden Markov chain and we can model its possible non stationarity by introducing an auxiliary random chain $U^2 = (U_1^2, ..., U_n^2)$, each U_i^2 taking its values in a finite set $\Lambda_2 = \{1, ..., M\}$. This leads to a TMC $T = (V, U^2, Y)$, which also is a TMC $T = (X, U, Y)$, with the auxiliary process $U = (U^1, U^2)$. Its distribution is given by $p(x, u^1, u^2)$ and $p(y | x, u^1, u^2) = p(y | x)$. Otherwise, the distribution $p(x, u^1, u^2)$ of (X, U^1, U^2) is given by $p(x_1, u_1^1, u_1^2)$ and the transitions $p(x_{i+1}, u_{i+1}^1, u_{i+1}^2 | x_i, u_i^1, u_i^2)$ that we can write as $p(x_{i+1}, u_{i+1}^1, u_{i+1}^2 | x_i, u_i^1, u_i^2) = p(u_{i+1}^2 | x_i, u_i^1, u_i^2) \times p(x_{i+1} | u_{i+1}^2, x_i, u_i^1, u_i^2) \times$ $\times p(u_{i+1}^1 | x_{i+1}, u_{i+1}^2 x_i, u_i^1, u_i^2)$. The particular transitions in this product, that define the new model we propose and that generalize formulas (1)-(2) above, are following:

$$p(u_{i+1}^2 | x_i, u_i^1, u_i^2) = \delta_{u_i^2}(u_{i+1}^2) \text{ if } u_i^1 > 1 \text{, and } p(u_{i+1}^2 | u_i^2) \text{ if } u_i^1 = 1 \text{ ;} \quad (3)$$

$$p(x_{i+1} | u_{i+1}^2, x_i, u_i^1, u_i^2) = \delta_{x_i}(x_{i+1}) \text{ if } u_i^1 > 1 \text{, and } p(x_{i+1} | u_{i+1}^2, x_i) \text{ if } u_i^1 = 1 \text{ ;} \quad (4)$$

$$p(u_{i+1}^1 | x_{i+1}, u_{i+1}^2 x_i, u_i^1, u_i^2) = \delta_{u_i^1 - 1}(u_{i+1}^1) \text{ if } u_i^1 > 1 \text{, and } p(u_{i+1}^1 | x_{i+1}, u_{i+1}^2) \text{ if } u_i^1 = 1 \text{ ;} \quad (5)$$

$p(x, u^1, u^2)$ being defined with $p(x_1, u_1^1, u_1^2)$ and (3)-(5), we end the definition of the distribution of $T = (X, U, Y)$ by considering $p(y | x, u^1, u^2) = p(y_1 | x_1) ... p(y_n | x_n)$.

Finally, putting $W = (X, U^1, U^2)$, we can say that $T = (W, Y)$ is a classical hidden Markov chain in which W is discrete and Y continuous. However, let us remark that the model is a particular one; in fact, we have $p(y_i | x_i, u_i^1, u_i^2) = p(y_i | x_i)$, which means that the noise distribution $p(y_i | x_i)$ neither depends on remaining sojourn time u_i^1, nor stationarity u_i^2. Of course, one can image that this noise distribution does depend on u_i^1, u_i^2, or even both of them, and the possibility of taking this into account in the model provide its possible further extensions.

Finally, having a classical hidden Markov chain allows us to compute $p(x_i, u_i^1, u_i^2 | y)$ by the classical use of "forward" and "backward" probabilities, which gives $p(x_i | y) = \sum_{u_i^1, u_i^2} p(x_i, u_i^1, u_i^2 | y)$. Used in the Bayesian MPM classification. Concerning the parameters estimation method we use the "Iterative Conditional Estimation" (ICE) described below.

PARAMETERS' ESTIMATION AND BAYESIAN SEGMENTATION

In experiments below we will use the following particular case of the model (3)-(5). We will consider that U_i^1 takes its values in a finite set $\Lambda_1 = \{1, ..., P\}$, and that for $u_i^1 = 1$ the probability $p(x_{i+1} = x_i | u_{i+1}^2, x_i)$ is not necessarily null. This condition means that for $U_i^1 = 1$ the value $U_{i+1}^1 = u_{i+1}^1$ is not the exact duration of sojourn in x_{i+1}, but the minimal duration. This defines a particular distribution of sojourn time on $N*$, which allows one to perform direct calculations, without resorting on Monte Carlo methods. Let us remark that a given distribution of W does not necessarily define an unique distribution of X; however, this problem does not arise in our experiments and we will not deal with any more in this paper.

From now, we will de by W the hidden process, which will be either $W = (X, U^1)$, or $W = (X, U^1, U^2)$. From the definition of the model seen above, W is a Markov chain and (W, Y) is a classical hidden Markov chain (HMC). We will assume that $p(y_i | x_i = \omega_j)$ does not depend on $i = 1, ..., n$ and is Gaussian with mean m_j and variance σ_j^2. Moreover, $p(w_i, w_{i+1})$ does not depend on $i = 1, ..., n-1$.

Finally, knowing that each (X_i, U_i^1, U_i^2) takes its values in a finite set $\{\omega_1, ..., \omega_K\} \times \{1, ..., M\} \times \{1, ..., P\}$, the whole model is defined for $W = (X, U^1, U^2)$, by $(KMP)^2$ real parameters giving the distribution $p(w_1, w_2)$, K means, and K variances. We propose to estimate all these parameters from $Y = y$ by a method derived from the general "Iterative Conditional estimation" (ICE).

According to its general principle, one can apply ICE to estimate a vector of parameters θ from Y once:

(i) there exists an estimator $\hat{\theta}(W, Y)$ of θ from complete data (W, Y);

(ii) for every θ, one can sample W according to $p(w | y, \theta)$.

The iterative ICE method runs as follows:

(i) consider an initial value θ^0;

(ii) put $\theta_r^{q+1} = E[\hat{\theta}_r(W, Y) | Y = y, \theta^q]$ for the components θ_r of θ for which this expectation is computable ;

(iii) for other components, sample m values $w^{q,1}$, …, $w^{q,m}$ of W according to $p(w | y, \theta^q)$, and put $\theta_r^{q+1} = \dfrac{\hat{\theta}_r(w^{q,1}, y) + ... + \hat{\theta}_r(w^{q,m}, y)}{m}$.

Let us consider the case $W = (X, U^1)$, as the case $W = (X, U^1, U^2)$ can be dealt with in a similar way. There are $(KM)^2$ parameters $p_{ij} = p(w_1 = i, w_2 = j)$ (therefore w_1 and w_2 are both in $\{\omega_1, ..., \omega_K\} \times \{1, ..., M\}$), K means m_1, …, m_K and K variances σ_1^2, …, σ_K^2. Denoting by I the indicator function, the classical estimator from complete data used is

$$\hat{p}_{ij}(w,y) = \frac{1}{n}\sum_{m=1}^{n} I(w_{2m-1} = i, w_{2m} = j) \tag{6}$$

$$\hat{m}_l(w,y) = \frac{1}{n}\sum_{m=1}^{n} y_l I(x_m = \omega_l) \tag{7}$$

$$\hat{\sigma}_l^2(w,y) = \frac{1}{n}\sum_{m=1}^{n} (y_l - \hat{m}_l(w,y))^2 I(x_m = \omega_l) \tag{8}$$

Recalling that the expectation of an incator function is the probability of the corresponding set and applying the conditional expectation $E[\hat{p}_{ij}(W,Y)\big|Y = y, \theta^q]$ to (6) gives

$$p_{ij}^{q+1}(y) = \frac{1}{m}\sum_{m=1}^{n} p(w_{2m-1} = i, w_{2m} = j\big|y, \theta^q), \tag{9}$$

while its application to (7) and (8) is not computable and we resort on sampling. This sampling is workable, as $p(w\big|y,\theta^q)$ is a Markov chain distribution with calculable transitions $p(w_{k+1}\big|w_k, y, \theta^q)$ (see below). Then we simulate one sample w^q (we take $m = 1$ in (iii)) and (7), (8) are applied to (w^q, y) instead of (w, y).
Finally, to perform unsupervised segmentation using ICE, we have to calculate the following three distributions: $p(w_{k+1}\big|w_k, y, \theta^q)$, $p(w_k, w_{k+1}\big|y, \theta^q)$ needed in ICE, and $p(w_k\big|y, \theta^q)$ needed in Bayesian MPM segmentation method. These distributions are classically computed from "forward" $\alpha_k(w_k) = p(w_k\big|y_1,...,y_k)$ and "backward" $\beta_k(w_k) = p(y_{k+1},...,y_n\big|w_k, y_k)$ probabilities, which are computed by the following forward (10) and backward (11) recursions

$$\alpha_1(w_1) = p(w_1), \text{ and } \alpha_{k+1}(w_{k+1}) = \sum_{w_k} \alpha_k(w_k) p(w_{k+1}, y_{k+1}\big|w_k, y_k) \text{ for } 2 \le k \le n-1; \tag{10}$$

$$\beta_n(w_n) = 1, \text{ and } \beta_k(w_k) = \sum_{w_{k+1}} \beta_{k+1}(w_{k+1}) p(w_{k+1}, y_{k+1}\big|w_k, y_k) \text{ for } 1 \le k \le n-1; \tag{11}$$

Then we have

$$p(w_k, w_{k+1}, y\big|\theta^q) = \alpha_k(w_k) p(w_{k+1}, y_{k+1}\big|w_k, y_k)\beta_{k+1}(w_{k+1}), \tag{12}$$

which gives $p(w_{k+1}\big|w_k, y, \theta^q)$, $p(w_k, w_{k+1}\big|y, \theta^q)$, and $p(w_k\big|y, \theta^q)$.

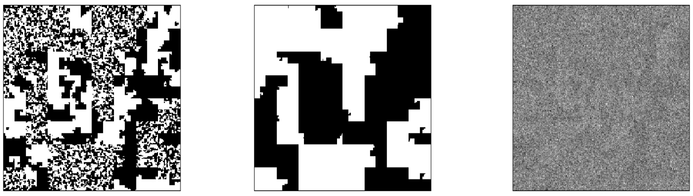
EXPERIMENTS

We present below two series of experiments.

In the first one we simulate a particular hidden semi-Markov non stationary chain, where X takes its values in $\Omega = \{\omega_1, \omega_2\}$, U^1 takes its values in $\Lambda_1 = \{1,...,5\}$, and U^2 takes its values in $\Lambda_2 = \{0,1\}$, which means that there are two different stationarities. The distributions $p(y_i|x_i)$ are normal a common standard deviation equal to 1, and the means equal to 1 for $x_i = \omega_1$, and equal 1.5 for $x_i = \omega_2$, respectively. We have $p(u_{i+1}^2|u_i^2) = \begin{pmatrix} 0.999 & 0.001 \\ 0.001 & 0.999 \end{pmatrix}$ in (3), $p(x_{i+1}|u_{i+1}^2 = 0, x_i) = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix}$ and $p(x_{i+1}|u_{i+1}^2 = 1, x_i) = \begin{pmatrix} 0.7 & 0.3 \\ 0.3 & 0.7 \end{pmatrix}$ in (4), and $p(u_{i+1}^1|x_{i+1}, u_{i+1}^2) = 0.2$ in (5). The observed $Y = y$ is then segmented by there unsupervised methods. The first method is based on the very classical HMC model, the second one is based on a stationary HSMC, and the last one is based on the proposed TMC, equivalent to a NSHSMC. Of course, as the data follow the new model, the very Bayesian theory requires that its use give best results. However, the experiment is of interest because the three segmentations are performed in unsupervised manner, and in a rather strongly noisy context. Then the theoretical superiority of NSHSMC based method is no longer true, and the use of a simpler model like HSMC, or even HMC, which contains less parameters to be estimated, could possibly produce better results than the use of NSHSMC. Let us notice that one possible application is image segmentation, where the use of mono-dimensional chains is possible by associating the mono-dimensional processes with the bi-dimensional set of pixels by using the Hilbert-Peano scan [1, 2, 6]. The images $X = x$, $U^2 = u^2$, and $Y = y$ so obtained are presented in Figure 1. The results show that the theoretic hierarchy is saved in the unsupervised segmentation: NSHSMC work better than stationary HSMC, and stationary HSMC work better than stationary HMC. Otherwise, in spite of the very high level of the noise (see $Y = y$ in Figure 1), their estimation with ICE give quite satisfying results when using NSHSMC (see Table1).

In the second experiments, we consider a two classes image $X = x$ and its noisy version $Y = y$ (see Figure 2, obtained by the use of Hilbert-Peano scan). As in the series above, $Y = y$ is segmented by the same three unsupervised methods. Of course, the data follow no one of the three models and thus the objective here is study how the three models work in such a case. As we can see in Table 2 and Figure 2, the same hierarchy is respected. Therefore we see that the NSHSMC based unsupervised method is better than the HSMC method, and the latter method is better than the HMC based one.

In ICE method used the initialization has been performed by the classical c-means method.



$X = x$ $U = u$ $Y = y$

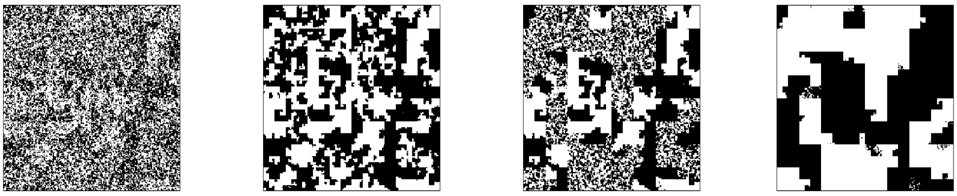
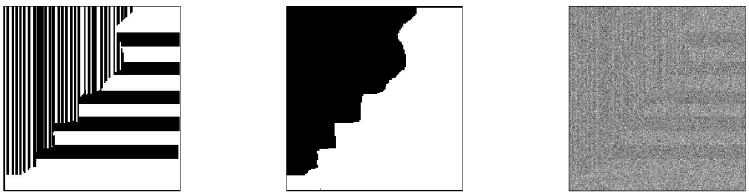


FIGURE 1. Second line, from left to right: segmentation of $Y = y$ with HMC (error ration : 34%), HSMC (error ration : 22%), and NSHSMC (error ration : 17%), estimation of U^2.

TABLE 1. Parameter's estimation using ICE

Classe	By HMC		By SemiHMC2		By SemiHMC-NS	
	Mean	Std deviation	Mean	Std deviation	Mean	Std deviation
0	0.8520	0.8871	1.06	1.02	1.0046	0.9899
1	1.661	0.9009	1.4602	1.0153	1.5137	0.9963
Error ratio	34%		22%		17%	



$X = x$ $U = u$ $Y = y$

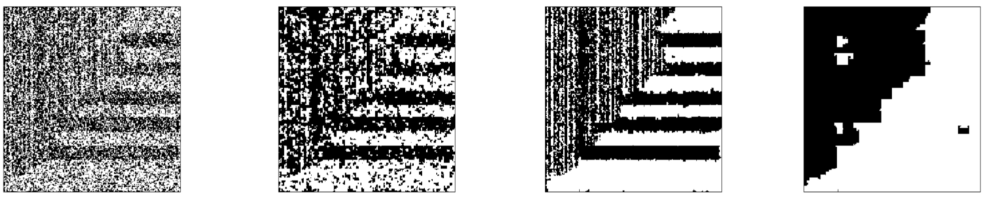


FIGURE 2. Second line, from left to right: segmentation of $Y = y$ with HMC (error ration : 35%), HSMC (error ration : 23%), and NSHSMC (error ration : 14%), estimation of U^2.

TABLE 2. Parameter's estimation using ICE

Classe	By HMC		By SemiHMC2		By SemiHMC-NS	
	Mean	Std deviation	Mean	Std deviation	Mean	Std deviation
0	0.84	0.91	1.09	1.04	0.9	0.94
1	1.65	0.89	1.46	1.02	1.49	0.99
Error ratio	35%		23%		14%	

CONCLUSION

In this paper, we have proposed a new model of a hidden non stationary semi-Markov chains. Extending some first suggestions presented in [9], the general idea was to use a triplet Markov chain $T=(X,U,Y)$ with $U=(U^1,U^2)$, where U^1 models the semi-markovianity, and U^2 models the non stationarity. As $T=(X,U,Y)$ is itself stationary, it is possible to estimate its parameters using the general "Iterative Conditional Estimation" (ICE) method, which leads to unsupervised Bayesian segmentation methods. We proposed two series of experiments which show that, on the one hand, the hidden semi-Markov chains based unsupervised segmentation method work better that the classical hidden Markov chains based unsupervised segmentation method and, on the other hand, the new model based unsupervised segmentation method work better that the hidden semi-Markov chains based one.

The classical hidden Markov chains are applied in various areas like Biosciences, Climatology, Communications, Ecology, Econometrics and Finance, Image or Signal processing. Therefore, the model we propose in this paper is likely to be useful and improve different processing in the same applications.

REFERENCES

1. B. Benmiloud, W. Pieczynski, Estimation des paramètres dans les chaînes de Markov cachées et segmentation d'images, Traitement du Signal, Vol. 12, No. 5, pp. 433-454, 1995.
2. S. Derrode and W. Pieczynski, Signal and Image Segmentation using Pairwise Markov Chains, IEEE Trans. on Signal Processing, Vol. 52, No. 9, pp. 2477-2489, 2004.
3. S. Faisan, L. Thoraval, J.-P. Armspach, M.-N. Metz-Lutz, and 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.
4. N. Giordana and W. Pieczynski, Estimation of Generalized Multisensor Hidden Markov Chains and Unsupervised Image, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 19, No. 5, pp. 465-475, 1997.
5. Y. Guédon, Estimating Hidden semi-Markov chains from discrete sequences, Journal of Computational and Graphical Statistics, Vol. 12, No. 3, pp. 604-639, 2003.
6. P. Lanchantin and W. Pieczynski, Unsupervised non stationary image segmentation using triplet Markov chains, Advanced Concepts for Intelligent Vision Systems (ACVIS 04), Aug. 31-Sept. 3, Brussels, Belgium, 2004.
7. 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.
8. 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.
9. W. Pieczynski and F. Desbouvries, On triplet Markov chains, International Symposium on Applied Stochastic Models and Data Analysis, (ASMDA 2005), Brest, France, May 2005.
10. 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.
11. S.-Z. Yu and H. Kobayashi, An efficient Forward-Backward algorithm for an explicit-duration hidden Markov model, IEEE Signal Processing Letters, Vol. 10, No. 1, pp. 11-14, 2003.