Parameter estimation in conditionally Gaussian pairwise Markov switching models and unsupervised smoothing

Fei Zheng, Stéphane Derrode, Wojciech Pieczynski

PARAMETER ESTIMATION IN CONDITIONALLY GAUSSIAN PAIRWISE MARKOV SWITCHING MODELS AND UNSUPERVISED SMOOTHING

Fei Zheng[a], *Stéphane Derrode*[a], *Wojciech Pieczynski*[b]

[a] École Centrale de Lyon, LIRIS, CNRS UMR 5205, Écully, France
[b] SAMOVAR, Telecom Sudparis, CNRS, Université Paris-Saclay, 9, rue Charles Fourier, Évry, France

ABSTRACT

Automatic identification of jump Markov systems (JMS) is known to be an important but difficult problem. In this work, we propose a new algorithm for the unsupervised estimation of parameters in a class of linear JMS called "conditionally Gaussian pairwise Markov switching models" (CGPMSMs), which extends the family of classic "conditionally Gaussian linear state-space models" (CGLSSMs). The method makes use of a particular CGPMSM called "conditionally Gaussian observed Markov switching model" (CGOMSM). The algorithm proposed consists in applying two EM algorithms sequentially: the first one is used to estimate the parameters and switches of the discrete pairwise Markov chain (PMC), which is a part of CGOMSM. Once estimated, it is used to sample switches and then the second one, called switching EM, is used to estimate the parameters of the distribution driving hidden states given the observations and the switches. The entire algorithm is evaluated with respect to data simulated according to CGPMSMs, and comparisons with several supervised methods attest its good efficiency.

Index Terms—Jump Markov linear systems, Expectation-Maximization, parameter estimation.

1. INTRODUCTION

Let us consider three random sequences $\mathbf{X}_1^N = (\mathbf{X}_1, \ldots, \mathbf{X}_N)$, $\mathbf{R}_1^N = (\mathrm{R}_1, \ldots, \mathrm{R}_N)$ and $\mathbf{Y}_1^N = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, taking their values in $\mathbb{R}^m$, $\boldsymbol{\Omega} = \{0, \ldots, K-1\}$ and $\mathbb{R}^q$, respectively.

$$\underbrace{\begin{bmatrix}\mathbf{x}_{n+1}\\ \mathbf{y}_{n+1}\end{bmatrix}}_{\mathbf{Z}_{n+1}} = \underbrace{\begin{bmatrix}\boldsymbol{\mathcal{F}}^{xx}(\mathbf{R}_n^{n+1}) & \boldsymbol{\mathcal{F}}^{xy}(\mathbf{R}_n^{n+1})\\ \boldsymbol{\mathcal{F}}^{yx}(\mathbf{R}_n^{n+1}) & \boldsymbol{\mathcal{F}}^{yy}(\mathbf{R}_n^{n+1})\end{bmatrix}}_{\boldsymbol{\mathcal{F}}(\mathbf{R}_n^{n+1})} \underbrace{\begin{bmatrix}\mathbf{x}_n\\ \mathbf{y}_n\end{bmatrix}}_{\mathbf{Z}_n} + \underbrace{\begin{bmatrix}\boldsymbol{\omega}_{n+1}^x\\ \boldsymbol{\omega}_{n+1}^y\end{bmatrix}}_{\boldsymbol{\omega}_{n+1}} + \underbrace{\begin{bmatrix}\mathbf{N}_{n+1}^x(\mathbf{R}_n^{n+1})\\ \mathbf{N}_{n+1}^y(\mathbf{R}_n^{n+1})\end{bmatrix}}_{\mathbf{N}_{n+1}^z(\mathbf{R}_n^{n+1})}. \quad (1)$$

$\mathbf{R}_1^N$ is a Markov chain and the couple $(\mathbf{X}_1^N, \mathbf{Y}_1^N)$ is assumed Markov and Gaussian conditionally on $\mathbf{R}_1^N$. The distribution of $(\mathbf{X}_1^N, \mathbf{R}_1^N, \mathbf{Y}_1^N)$ is defined by $p(r_1)$, Gaussian distributions $p(\mathbf{x}_1, \mathbf{y}_1 | r_1)$, transitions $p(r_{n+1} | r_n)$, and the system (1), in which $\boldsymbol{\omega_{n+1}}$ represents the noises independent of z_1 and

$$\boldsymbol{\omega_{n+1}} \sim \mathcal{N}\left(\mathbf{0}, \underbrace{\begin{bmatrix}\boldsymbol{\mathcal{Q}}^{xx}(\mathbf{R}_n^{n+1}) & \boldsymbol{\mathcal{Q}}^{xy}(\mathbf{R}_n^{n+1})\\ \boldsymbol{\mathcal{Q}}^{yx}(\mathbf{R}_n^{n+1}) & \boldsymbol{\mathcal{Q}}^{yy}(\mathbf{R}_n^{n+1})\end{bmatrix}}_{\boldsymbol{\mathcal{Q}}(\mathbf{R}_n^{n+1})}\right).$$

Thus, system parameters $\boldsymbol{\mathcal{F}}(\mathbf{R}_n^{n+1})$, $\boldsymbol{\mathcal{Q}}(\mathbf{R}_n^{n+1})$ depend on the switches $\mathbf{R}_n^{n+1} = (\mathrm{R}_n, \mathrm{R}_{n+1})$, and $\mathbf{N}_{n+1}^z(\mathbf{R}_n^{n+1})$ denotes the item about the mean values,

$$\underbrace{\begin{bmatrix}\mathbf{N}_{n+1}^x(\mathbf{R}_n^{n+1})\\ \mathbf{N}_{n+1}^y(\mathbf{R}_n^{n+1})\end{bmatrix}}_{\mathbf{N}_{n+1}^z(\mathbf{R}_n^{n+1})} = \underbrace{\begin{bmatrix}\mathbf{M}^x(\mathrm{R}_{n+1})\\ \mathbf{M}^y(\mathrm{R}_{n+1})\end{bmatrix}}_{\mathbf{M}^z(\mathrm{R}_{n+1})} - \boldsymbol{\mathcal{F}}(\mathbf{R}_n^{n+1}) \underbrace{\begin{bmatrix}\mathbf{M}^x(\mathrm{R}_n)\\ \mathbf{M}^y(\mathrm{R}_n)\end{bmatrix}}_{\mathbf{M}^z(\mathrm{R}_n)},$$

where $\mathbf{M}^x(\mathrm{R}_n)$ and $\mathbf{M}^y(\mathrm{R}_n)$ are the means of the two processes respectively, which are only decided by the value of R_n and without dependence on n.

Such a system, named "conditionally Gaussian pairwise Markov switching model" CGPMSM [1], extends the well-known "conditionally Gaussian linear state-space model" (CGLSSM) [2-3]. The latter in which $\boldsymbol{\mathcal{F}}^{xy}(\mathbf{R}_n^{n+1})$ and $\boldsymbol{\mathcal{F}}^{yy}(\mathbf{R}_n^{n+1})$ in (1) are set to be zero, is considered as the "natural" switching Gaussian system, however, its application is quite limited as it does not allow fast optimal filters [4-5]. Another recent particular case of CGPMSM, called "conditionally Gaussian observed Markov switching model" (CGOMSM), consists in taking CGPMSM with $\boldsymbol{\mathcal{F}}^{yx}(\mathrm{R}_{n+1}) = \mathbf{O}$. Decisive advantage of CGOMSM over classic switching systems is that it does allow fast exact optimal filtering.

This paper proposes a novel algorithm to automatically estimate the parameters $\{\boldsymbol{\mathcal{F}}(\mathbf{R}_n^{n+1}), \boldsymbol{\mathcal{Q}}(\mathbf{R}_n^{n+1}), \mathbf{M}^y(\mathrm{R}_{n+1})\}$ and the parameters defining the distribution of the switches of CGPMSM described above from a limited-size set of

observations only, with $\mathbf{M}^x(\mathrm{R}_{n+1})$ assumed to be known as zero. The interest of such a general algorithm relies on the fact that it is theoretically possible to approximate any non-linear and non-Gaussian system by the linear switching system (1).

The new proposed method, called "double expectation-maximization-CGPMSM" (DEM-CGPMSM), uses the well-known EM method at two different levels:

(i) The first one is to consider that $(\mathbf{R}_1^N, \mathbf{Y}_1^N)$ is Markov, and to use EM to estimate its parameters from $\mathbf{Y}_1^N$. This is an approximation, as $(\mathbf{R}_1^N, \mathbf{Y}_1^N)$ is not necessarily Markov in general CGPMSM. However, this approximation does not seem too strong for what is to be done, and different experiments show a good robustness of the whole unsupervised filtering method with respect to it;

(ii) Once the parameters of $(\mathbf{R}_1^N, \mathbf{Y}_1^N)$ are estimated, they are used to sample a realization $\mathbf{r}_1^N$ of $\mathbf{R}_1^N$ using $p(\mathbf{r}_1^N | \mathbf{y}_1^N)$, and taking $\mathbf{r}_1^N$ as it were a true realization of $\mathbf{R}_1^N$. Then the "true" switches so obtained are taken for applying an adaptation of a recent EM [8] for estimating the parameters of $p(\mathbf{x}_1^N, \mathbf{y}_1^N | \mathbf{r}_1^N)$.

The proposed method can be considered as an extension to the classical EM method [6-9] for switching cases, and also as an alternative to particle filter based methods [10-11], which can be time consuming. Moreover, it is more general than the ones described in [12], in which the system at the origin of the observations is supposed to be known (and in which sampling can be used for parameters estimation).

The remaining of the paper is organized as follows. Section 2 explains the implementation details of the DEM-CGPMSM algorithm proposed. The Switching EM and the integrated DEM-CGPMSM algorithm are tested in Section 3 to verify the robustness of parameter estimation and smoothing performance against several supervised methods. Some potential impact factors on the performance of our algorithm are analyzed. Finally, Section 4 presents further possible improvements for our algorithm.

2. DOUBLE EM ALGORITHM FOR PARAMETER ESTIMATION OF CGPMSM

As mentioned in the Introduction, the DEM-CGPMSM is based on two EM algorithms. It runs as follows:

1. Consider the distribution of $(\mathbf{R}_1^N, \mathbf{Y}_1^N)$ in CGPMSM as being a PMC distribution; apply EM to estimate the related parameters. Sample $\mathbf{R}_1^N = \mathbf{r}_1^N$ according to $p(\mathbf{r}_1^N | \mathbf{y}_1^N)$, and estimate $\mathbf{M}^y(\mathrm{R}_n)$ through empirical estimation;
2. Consider the $\mathbf{R}_1^N = \mathbf{r}_1^N$ sampled and the means $\mathbf{M}^y(\mathrm{R}_n)$ estimated in step 1. The second EM (Switching EM) is applied to get the remaining parameters $\mathcal{F}(\mathbf{R}_n^{n+1})$ and $\mathcal{Q}(\mathbf{R}_n^{n+1})$. Use the estimated parameters to restore $\mathbf{X}_1^N$.

2.1. EM for discrete pairwise Markov chain

Let us consider $\mathbf{R}_1^N = (\mathrm{R}_1, \ldots, \mathrm{R}_N)$, $\mathbf{Y}_1^N = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$ above, and $\mathbf{T}_1^N = (\mathbf{R}_1^N, \mathbf{Y}_1^N)$, with $\mathbf{T}_n = (\mathrm{R}_n, \mathbf{Y}_n)$. The process $\mathbf{T}_1^N$ is called a "pairwise Markov chain" (PMC), if it is Markov, i.e. its distribution can be written as:

$$p(\mathbf{t}_1^N) = p(\mathbf{t}_1)\, p(\mathbf{t}_2 | \mathbf{t}_1) \ldots p(\mathbf{t}_N | \mathbf{t}_{N-1}). \quad (2)$$

Assuming that the PMC is "Stationary" and "Gaussian", which means that the $p(\mathbf{t}_n | \mathbf{t}_{n+1})$ do not depend on n and $p(\mathbf{y}_n, \mathbf{y}_{n+1} | r_n, r_{n+1})$ are Gaussian. The distribution is then given by:

$$p(\mathbf{t}_1, \mathbf{t}_2) = p(j,k)\, f_{j,k}(\mathbf{y}_1, \mathbf{y}_2), \quad (3)$$

where

$$p(j,k) = p(r_n = j, r_{n+1} = k);$$
$$f_{j,k}(\mathbf{y}_1, \mathbf{y}_2) = p(\mathbf{y}_1, \mathbf{y}_2 | r_n = j, r_{n+1} = k) \sim \mathcal{N}(\boldsymbol{\mu}_y^{j,k}, \boldsymbol{\Gamma}_y^{j,k}).$$

with $j, k \in \{0, K-1\}$. EM can be applied on this Gaussian PMC system [13]. For later use, we define:

$$\begin{cases} \phi_n(j) = p(r_n = j | \mathbf{y}_1^N) \\ \psi_n(j,k) = p(r_n = j, r_{n+1} = k | \mathbf{y}_1^N) \end{cases}. \quad (4)$$

2.1.1. Forward & Backward probabilities

$\phi_n(j)$ and $\psi_n(j,k)$ in (4) can be calculated recursively by Baum's algorithm [14]:

$$\phi_n(j) = \alpha_n(j)\, \beta_n(j); \quad (5)$$

$$\psi_n(j,k) = \frac{\alpha_n(j)\, p(r_{n+1} = k, \mathbf{y}_{n+1} | r_n = j, \mathbf{y}_n)\, \beta_{n+1}(k)}{\sum_{(l_1,l_2)\in\Omega^2} \alpha_n(l_1)\, p(r_{n+1} = l_2, \mathbf{y}_{n+1} | r_n = l_1, \mathbf{y}_n)\, \beta_{n+1}(l_2)}, \quad (6)$$

where α_n and β_n denote the "normalized" forward and backward probabilities defined by:

$$\alpha_n(j) = p(r_n = j | \mathbf{y}_1^n), \quad (7)$$

$$\beta_n(j) = \frac{p(\mathbf{y}_{n+1}^N | r_n = j, \mathbf{y}_n)}{p(\mathbf{y}_{n+1}^N | \mathbf{y}_1^n)}. \quad (8)$$

The recursive computation of $\alpha_n(j)$ and $\beta_n(j)$ are:

$$\begin{cases} \alpha_1(j) = \dfrac{p(r_1 = j, \mathbf{y}_1)}{\sum_{l\in\Omega} p(r_1 = l, \mathbf{y}_1)} \\ \alpha_n(j) = \dfrac{\sum_{l\in\Omega} \alpha_{n-1}(l)\, p(r_n = j, \mathbf{y}_n | r_{n-1} = l, \mathbf{y}_{n-1})}{\sum_{(l_1,l_2)\in\Omega^2} \alpha_{n-1}(l_1)\, p(r_n = l_2, \mathbf{y}_n | r_{n-1} = l_1, \mathbf{y}_{n-1})} \\ \beta_N(j) = 1 \\ \beta_n(j) = \dfrac{\sum_{l\in\Omega} \beta_{n+1}(l)\, p(r_{n+1} = l, \mathbf{y}_{n+1} | r_n = j, \mathbf{y}_n)}{\sum_{(l_1,l_2)\in\Omega^2} \alpha_n(l_1)\, p(r_{n+1} = l_2, \mathbf{y}_{n+1} | r_n = l_1, \mathbf{y}_n)} \end{cases}. \quad (9)$$

2.1.2. EM parameter estimation

For starting the EM, the switches are initialized by K-means. $\delta_n^{n+1}(j,k)$ is denoted as function $\mathbb{1}_{[r_n=j,r_{n+1}=k]}$ and $\boldsymbol{Card}(j,k)=\sum_{n=1}^{N-1}\delta_n^{n+1}(j,k)$. The initialization of parameters is calculated from the following empirical equations:

$$\begin{cases}\hat{p}(j,k)=\dfrac{\boldsymbol{Card}(j,k)}{N-1}\\ \hat{\boldsymbol{\mu}}_y^{j,k}=\dfrac{1}{\boldsymbol{Card}(j,k)}\sum_{n=1}^{N-1}\delta_n^{n+1}(j,k)\left(\begin{bmatrix}\mathbf{y}_n\\ \mathbf{y}_{n+1}\end{bmatrix}\right)\\ \hat{\boldsymbol{\Gamma}}_y^{j,k}=\dfrac{1}{\boldsymbol{Card}(j,k)}\sum_{n=1}^{N-1}\delta_n^{n+1}(j,k)\left(\begin{bmatrix}\mathbf{y}_n\\ \mathbf{y}_{n+1}\end{bmatrix}-\hat{\boldsymbol{\mu}}_y^{j,k}\right)\left(\begin{bmatrix}\mathbf{y}_n\\ \mathbf{y}_{n+1}\end{bmatrix}-\hat{\boldsymbol{\mu}}_y^{j,k}\right)^t\end{cases} \tag{10}$$

and the EM updates the parameters by maximizing the likelihood function of $E\left[\ln p\left(\mathbf{T}_1^N\right)|\mathbf{y}_1^N\right]$, as

$$\begin{cases}\hat{p}(j,k)=\dfrac{1}{N-1}\sum_{n=1}^{N-1}\psi_n(j,k)\\ \hat{\boldsymbol{\mu}}_y^{j,k}=\dfrac{\sum_{n=1}^{N-1}\psi_n(j,k)\left(\begin{bmatrix}\mathbf{y}_n\\ \mathbf{y}_{n+1}\end{bmatrix}\right)}{\sum_{n=1}^{N-1}\psi_n(j,k)}\\ \hat{\boldsymbol{\Gamma}}_y^{j,k}=\dfrac{\sum_{n=1}^{N-1}\psi_n(j,k)\left(\begin{bmatrix}\mathbf{y}_n\\ \mathbf{y}_{n+1}\end{bmatrix}-\hat{\boldsymbol{\mu}}_y^{j,k}\right)\left(\begin{bmatrix}\mathbf{y}_n\\ \mathbf{y}_{n+1}\end{bmatrix}-\hat{\boldsymbol{\mu}}_y^{j,k}\right)^t}{\sum_{n=1}^{N-1}\psi_n(j,k)}\end{cases}. \tag{11}$$

Iterations are stopped after sufficient iterations, assuming that the estimated parameters do not change a lot. Once $\phi_n(j)$ are got, Maximizer of the posterior marginal (MPM) is applied for sampling $\mathbf{r}_1^N$ with $p\left(\mathbf{r}_1^N\left|\mathbf{y}_1^N\right.\right)$, then $\mathbf{M}^y(\mathrm{R}_n)$ are estimated through the empirical estimation:

$$\hat{\mathbf{M}}^y(\mathrm{R}_n=j)=\hat{\mathbf{M}}^y(j)=\frac{\sum_1^N\mathbf{Y}_n(r_n=j)}{\boldsymbol{Card}(j)}, \tag{12}$$

in which

$$\boldsymbol{Card}(j)=\textstyle\sum_1^N\mathbb{1}_{[r_n=j]}.$$

2.2. Switching EM

This section is based on the assumption that switches are known, and the object is to expand the recent EM for parameter estimation of pairwise Markov models to a "Switching EM", which is new suiting the switch models.

2.2.1. Forward & Backward probabilities

Assuming the switches $\mathbf{R}_1^N=\mathbf{r}_1^N$ are given, from the properties of CGPMSM, we have:

$$\begin{aligned}&p\left(\mathbf{x}_{n+1},\mathbf{y}_{n+1}\left|\mathbf{x}_n,\mathbf{r}_n^{n+1},\mathbf{y}_n\right.\right)\\ &\sim\mathcal{N}\left(\boldsymbol{\mathcal{F}}(\mathbf{r}_n^{n+1})\mathbf{z}_n+\mathbf{N}_{n+1}^z(\mathbf{r}_n^{n+1}),\boldsymbol{\mathcal{Q}}(\mathbf{r}_n^{n+1})\right).\end{aligned} \tag{13}$$

To simplify, let us remove the fixed $\mathbf{r}_n^{n+1}$ in the derivation. Calculate the intermediates for forward probabilities:

$$\hat{\mathbf{x}}_{n|n+1}=\hat{\mathbf{x}}_{n|n}+\boldsymbol{K}_{n|n+1}\tilde{\mathbf{y}}_{n+1|n}; \tag{14}$$

$$\boldsymbol{P}_{n|n+1}=\boldsymbol{P}_{n|n}-\boldsymbol{K}_{n|n+1}\boldsymbol{S}_{n|n+1}\left(\boldsymbol{K}_{n|n+1}\right)^t, \tag{15}$$

with

$$\begin{aligned}\boldsymbol{S}_{n|n+1}&=\boldsymbol{\mathcal{Q}}^{yy}+\boldsymbol{\mathcal{F}}^{yx}\boldsymbol{P}_{n|n}\left(\boldsymbol{\mathcal{F}}^{yx}\right)^t;\\ \boldsymbol{K}_{n|n+1}&=\boldsymbol{P}_{n|n}\left(\boldsymbol{\mathcal{F}}^{yx}\right)^t\left(\boldsymbol{S}_{n|n+1}\right)^{-1};\\ \hat{\mathbf{y}}_{n+1|n}&=\boldsymbol{\mathcal{F}}^{yx}\hat{\mathbf{x}}_{n|n}+\boldsymbol{\mathcal{F}}^{yy}\mathbf{y}_n+\mathbf{N}_{n+1}^y;\\ \tilde{\mathbf{y}}_{n+1|n}&=\mathbf{y}_{n+1}-\hat{\mathbf{y}}_{n+1|n},\end{aligned}$$

therefore,

$$\hat{\mathbf{x}}_{n+1|n+1}=\boldsymbol{A}_n\hat{\mathbf{x}}_{n|n+1}+\boldsymbol{B}_n; \tag{16}$$

$$\boldsymbol{P}_{n+1|n+1}=\boldsymbol{\mathcal{Q}}_2+\boldsymbol{A}_n\boldsymbol{P}_{n|n+1}\left(\boldsymbol{A}_n\right)^t, \tag{17}$$

where

$$\begin{aligned}\boldsymbol{A}_n&=\boldsymbol{\mathcal{F}}^{xx}-\boldsymbol{\mathcal{Q}}^{xy}\left(\boldsymbol{\mathcal{Q}}^{yy}\right)^{-1}\boldsymbol{\mathcal{F}}^{yx};\\ \boldsymbol{B}_n&=\boldsymbol{\mathcal{Q}}^{xy}\left(\boldsymbol{\mathcal{Q}}^{yy}\right)^{-1}\mathbf{y}_{n+1}-\boldsymbol{\mathcal{Q}}^{xy}\left(\boldsymbol{\mathcal{Q}}^{yy}\right)^{-1}\mathbf{N}_{n+1}^y\\ &\quad+\left(\boldsymbol{\mathcal{F}}^{xy}-\boldsymbol{\mathcal{Q}}^{xy}\left(\boldsymbol{\mathcal{Q}}^{yy}\right)^{-1}\boldsymbol{\mathcal{F}}^{yy}\right)\mathbf{y}_n+\mathbf{N}_{n+1}^x;\\ \boldsymbol{\mathcal{Q}}_2&=\boldsymbol{\mathcal{Q}}^{xx}-\boldsymbol{\mathcal{Q}}^{xy}\left(\boldsymbol{\mathcal{Q}}^{yy}\right)^{-1}\boldsymbol{\mathcal{Q}}^{yx}.\end{aligned}$$

Backward probabilities (smoothing):

$$\hat{\mathbf{x}}_{n|N}=\hat{\mathbf{x}}_{n|n+1}+\boldsymbol{K}_{n|N}(\hat{\mathbf{x}}_{n+1|N}-\hat{\mathbf{x}}_{n+1|n+1}); \tag{18}$$

$$\boldsymbol{P}_{n|N}=\boldsymbol{P}_{n|n+1}+\boldsymbol{K}_{n|N}(\boldsymbol{P}_{n+1|N}-\boldsymbol{P}_{n+1|n+1})\left(\boldsymbol{K}_{n|N}\right)^t; \tag{19}$$

$$\boldsymbol{C}_{n+1,n|N}=\boldsymbol{P}_{n+1|N}\left(\boldsymbol{K}_{n|N}\right)^t, \tag{20}$$

in which

$$\boldsymbol{K}_{n|N}=\boldsymbol{P}_{n|n+1}\left(\boldsymbol{A}_n\right)^t\left(\boldsymbol{P}_{n+1|n+1}\right)^{-1}.$$

2.2.2. EM algorithm for switching model

Let $\Theta=(\Theta_1,\underbrace{\Theta_2^{1,1},\Theta_2^{1,2},\ldots,\Theta_2^{j,k},\ldots,\Theta_2^{K,K}}_{\Theta_2})$ be the parameters in CGPMSM, in which $\Theta_1=(\hat{\mathbf{x}}_1,\boldsymbol{P}_1)$, assuming that $\mathbf{x}_1\sim\mathcal{N}(\hat{\mathbf{x}}_1,\boldsymbol{P}_1)$; $\Theta_2^{j,k}=(\boldsymbol{\mathcal{F}}^{j,k},\boldsymbol{\mathcal{Q}}^{j,k})$, with $\boldsymbol{\mathcal{F}}^{j,k}=\boldsymbol{\mathcal{F}}\left(\mathbf{r}_n^{n+1}=(j,k)\right)$ and $\boldsymbol{\mathcal{Q}}^{j,k}=\boldsymbol{\mathcal{Q}}\left(\mathbf{r}_n^{n+1}=(j,k)\right)$. The recursion function of the EM algorithm is

$$\Theta^{(ii)}=arg\max_{\Theta}L(\Theta^{(ii-1)},\Theta), \tag{21}$$

where

$$L(\Theta^{(ii-1)},\Theta)=\mathrm{E}_{\Theta^{(ii-1)}}\left[\ln p_\Theta\left(\mathbf{z}_1^N\right)\left|\mathbf{y}_1^N\right.\right].$$

Let

$$\begin{cases}\mathbf{z}'_{n+1}=\mathbf{z}_{n+1}-\mathbf{M}^z(r_{n+1})\\ \mathbf{z}'_n=\mathbf{z}_n-\mathbf{M}^z(r_n)\end{cases},$$

$$\begin{aligned}\boldsymbol{C}_{z'_n,z'_n}&=E(\mathbf{z}'_n\mathbf{z}'^t_n|\mathbf{y}_1^N)\\ &=\begin{bmatrix}\hat{\mathbf{x}}_{n|N}-\mathbf{M}^x(r_n)\\ \mathbf{y}_n-\mathbf{M}^y(r_n)\end{bmatrix}\begin{bmatrix}\hat{\mathbf{x}}_{n|N}-\mathbf{M}^x(r_n)\\ \mathbf{y}_n-\mathbf{M}^y(r_n)\end{bmatrix}^t+\begin{bmatrix}\boldsymbol{P}_{n|N}&0\\ 0&0\end{bmatrix}.\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{C}_{z'_{n+1},z'_n}&=E(\mathbf{z}'_{n+1}\mathbf{z}'^t_n|\mathbf{y}_1^N)\\ &=\begin{bmatrix}\hat{\mathbf{x}}_{n+1|N}-\mathbf{M}^x(r_{n+1})\\ \mathbf{y}_{n+1}-\mathbf{M}^y(r_{n+1})\end{bmatrix}\begin{bmatrix}\hat{\mathbf{x}}_{n|N}-\mathbf{M}^x(r_n)\\ \mathbf{y}_n-\mathbf{M}^y(r_n)\end{bmatrix}^t+\begin{bmatrix}\boldsymbol{C}_{n+1,n|N}&0\\ 0&0\end{bmatrix}.\end{aligned}$$

Maximizing the likelihood function of the complete data, we can get the recursive EM update for $\Theta_2^{j,k}$ and Θ_1.

$$\boldsymbol{\mathcal{F}}^{j,k} = \widetilde{\boldsymbol{C}}^{j,k}_{z'_{n+1},z'_n}\left(\widetilde{\boldsymbol{C}}^{j,k}_{z'_n,z'_n}\right)^{-1}; \tag{22}$$

$$\boldsymbol{\mathcal{Q}}^{j,k} = \frac{1}{\boldsymbol{Card}\,(j,k)}\left(\widetilde{\boldsymbol{C}}^{j,k}_{z'_{n+1},z'_{n+1}} - \widetilde{\boldsymbol{C}}^{j,k}_{z'_{n+1},z'_n}\left(\widetilde{\boldsymbol{C}}^{j,k}_{z'_n,z'_n}\right)^{-1}\left(\widetilde{\boldsymbol{C}}^{j,k}_{z'_{n+1},z'_n}\right)^t\right), \tag{23}$$

where

$$\widetilde{\boldsymbol{C}}^{j,k}_{z'_n,z'_n} = \sum_{n=1}^{N-1}\delta_n^{n+1}(j,k)\,\boldsymbol{C}_{z'_n,z'_n};$$

$$\widetilde{\boldsymbol{C}}^{j,k}_{z'_{n+1},z'_n} = \sum_{n=1}^{N-1}\delta_n^{n+1}(j,k)\,\boldsymbol{C}_{z'_{n+1},z'_n};$$

$$\widetilde{\boldsymbol{C}}^{j,k}_{z'_{n+1},z'_{n+1}} = \sum_{n=1}^{N-1}\delta_n^{n+1}(j,k)\,\boldsymbol{C}_{z'_{n+1},z'_{n+1}},$$

and the update of $\hat{\mathbf{x}}_1$ and $\boldsymbol{P}_1$ are:

$$\hat{\mathbf{x}}_1 = \hat{\mathbf{x}}_{1|N}, \quad \boldsymbol{P}_1 = \hat{\boldsymbol{P}}_{1|N}. \tag{24}$$

2.3. Integrated Double EM-CGPMSM algorithm

The integrated DEM-CGPMSM algorithm is constructed by the EM for discrete PMC and the Switching EM described in previous subsections. For reducing the failure of parameter initialization, a feedback of parameters: $\mathbf{M}^y, \boldsymbol{\mathcal{F}}^{j,k}$ and $\boldsymbol{\mathcal{Q}}^{j,k}$ is added from the result of Switching EM to EM for discrete PMC. It is based on the assumption that data follows a homogeneous CGPMSM system, in which the parameters do not depend on n, given by the covariance matrix of $\mathbf{Z}_1^2 = [\mathbf{Z}_1(\mathrm{R}_1 = j), \mathbf{Z}_2(\mathrm{R}_2 = k)]$ as

$$\boldsymbol{\Gamma}^{j,k}_{\mathbf{Z}_1^2} = \begin{bmatrix} \boldsymbol{\Gamma}_z^j & \boldsymbol{\Sigma}_z^{j,k} \\ \left(\boldsymbol{\Sigma}_z^{j,k}\right)^t & \boldsymbol{\Gamma}_z^k \end{bmatrix}, \tag{25}$$

with

$$\boldsymbol{\Gamma}_z^l = \begin{bmatrix} \left(\boldsymbol{\sigma}_x^l\right)^2 & \boldsymbol{b}^l \\ \left(\boldsymbol{b}^l\right)^t & \left(\boldsymbol{\sigma}_y^l\right)^2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_z^{j,k} = \begin{bmatrix} \boldsymbol{a}^{j,k} & \boldsymbol{d}^{j,k} \\ \boldsymbol{e}^{j,k} & \boldsymbol{c}^{j,k} \end{bmatrix},$$

for all $l \in \{0, K-1\}$. Thus, the conversion formula between $\boldsymbol{\mathcal{F}}^{j,k}$, $\boldsymbol{\mathcal{Q}}^{j,k}$ and the co-variances are given by:

$$\boldsymbol{\mathcal{F}}^{j,k} = \left(\boldsymbol{\Sigma}_z^{j,k}\right)^t\left(\boldsymbol{\Gamma}_z^j\right)^{-1}; \quad \boldsymbol{\mathcal{Q}}^{j,k} = \boldsymbol{\Gamma}_z^k - \boldsymbol{\mathcal{F}}^{j,k}\boldsymbol{\Sigma}_z^{j,k}. \tag{26}$$

The feedback parameters are calculated through the inverse of relations in (26) and empirical estimation from parameters estimated by the two EMs at previous iteration. The $\boldsymbol{\Gamma}^{j,k}_{\mathbf{Z}_1^2}$ can be reconstructed as:

$$\hat{\boldsymbol{\Gamma}}_z^l = argvec\left[\left(\mathbf{I} - \hat{\boldsymbol{\mathcal{F}}}^{l,l} \otimes \hat{\boldsymbol{\mathcal{F}}}^{l,l}\right)^{-1} vec\left(\hat{\boldsymbol{\mathcal{Q}}}^{l,l}\right)\right]; \tag{27}$$

$$\hat{\boldsymbol{\Sigma}}_z^{j,k} = \left(\boldsymbol{\mathcal{F}}^{j,k}\hat{\boldsymbol{\Gamma}}_z^j\right)^t,$$

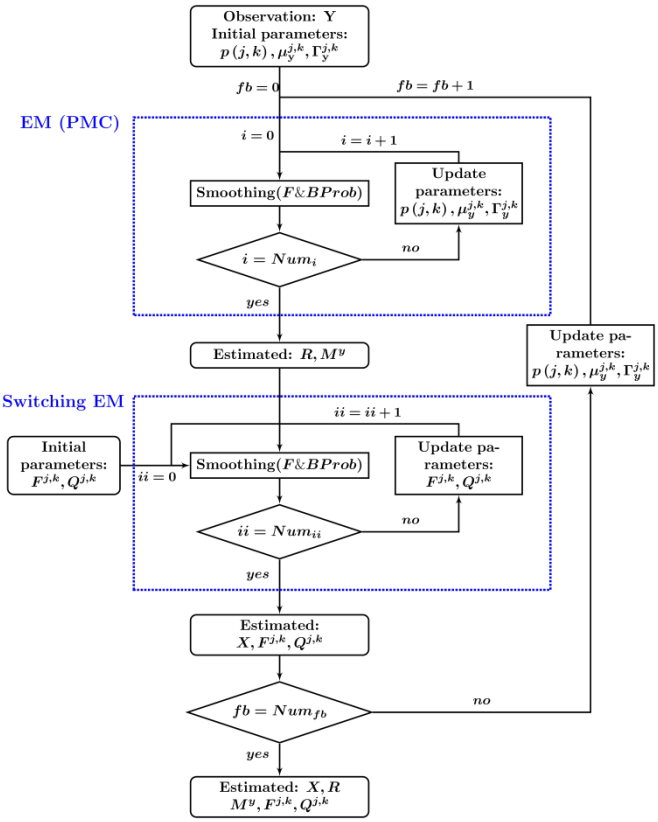


Fig. 1. Flowchart of DEM-CGPMSM.

in which $argvec(.)$ is the inverse function of the operator vector that stacks the columns of a matrix and $\otimes$ represents the Kronecker product. So, the feedback $\hat{\boldsymbol{\Gamma}}_y^{j,k}$ can be restored by the corresponding covariance elements of $[\mathbf{Y}_1, \mathbf{Y}_2]$ in $\boldsymbol{\Gamma}^{j,k}_{\mathbf{Z}_1^2}$ estimated. The other two feedback parameters are calculated as:

$$\hat{p}(j,k) = \frac{\boldsymbol{Card}\,(j,k)}{N-1}; \quad \hat{\boldsymbol{\mu}}_y^{j,k} = \left[\hat{\mathbf{M}}^y(j) \quad \hat{\mathbf{M}}^y(k)\right]^t. \tag{28}$$

The flowchart of the integrated DM-CGPMSM algorithm is shown in Fig. 1.

3. EXPERIMENTS

We present here the results of two series of experiments on scalar data ($m = q = 1$). The first series allows analyzing the robustness of the switching EM part only, and the second series experiments the global DEM-CGPMSM algorithm. Evaluations are performed using the Mean Square Error (MSE) between the estimated states and the simulated ones (ground truth). Comparisons are performed with respect to the Optimal Smoother Approximation (OSA) and other supervised methods which assume that all parameters are known.

For simplification, all experiments assume two jumps only and $\boldsymbol{\mathcal{F}}(\mathbf{R}_n^{n+1}) = \boldsymbol{\mathcal{F}}(\mathrm{R}_{n+1})$, $\boldsymbol{\mathcal{Q}}(\mathbf{R}_n^{n+1}) = \boldsymbol{\mathcal{Q}}(\mathrm{R}_{n+1})$, with transition matrix: $p(r_n = 0) = p(r_n = 1) = 0.5$ and $p(r_{n+1} = 0 \,|r_n = 0) = p(r_{n+1} = 1 \,|r_n = 1) = 0.9$.

The abbreviations and meaning of all comparative method used in this section are:

1. DEM (no feedback): DEM-CGPMSM without feedback;
2. DEM (one feedback): DEM-CGPMSM with one feedback;
3. OS: Optimal Smoothing with true $\mathbf{R}_1^N = \mathbf{r}_1^N$ and true parameters
4. OSA: Optimal Smoothing Approximation with true parameters modified to be CGOMSM.
5. CGLSSM: Classical Smoothing with true $\mathbf{R}_1^N = \mathbf{r}_1^N$ and true parameters modified to be CGLSSM.

3.1. Experiment regarding the Switching EM

The aim of this series is to test the performance of the Switching EM which is a part of our integrated method for parameter estimation. N = 2000 samples of $(\mathbf{X}_1^N, \mathbf{R}_1^N, \mathbf{Y}_1^N)$ are simulated according to a CGPMSM with zero means.

The true parameters for data simulation are reported in column "Θ" of Table II, then parameters are estimated from the observation and switches through Switching EM with 500 iterations. The initialization was set according to column "$\Theta^{(0)}$" in which "$Var[Y(R)]$" means to take the variance of the subseries of $\mathbf{Y}_n$ according to current value of R_n. The same initialization and iteration are executed when doing parameter estimation through Classical EM which assumes no switch in data. The parameters estimated from Switching EM and Classical EM are shown respectively under the column "$\Theta^{(500)}_{SwitchingEM}$" and column "$\Theta^{(500)}_{ClassicalEM}$".

Table I shows all the restoration MSE including the Optimal Kalman Smoothing (or Optimal Smoothing) with true parameters for comparison. Results are averages of 100 independent experiments.

From table I, we find that Switching EM is more suitable for switch models than Classical EM; in addition, its performance can be close to the Optimal Smoothing with true parameters. Besides, the parameters estimated through Switching EM can be very similar to the true parameters, even with initial values set far from the true ones.

3.2. Experiment regarding the entire DEM-CGPMSM

The aim of this series is to analyze the performances of DEM-CGPMSM with respect to several supervised methods.

Data is sampled according to a homogeneous CGPMSM described as (25). To simplify, we assume that all variances are 1 and $b^0 = b^1 = b$. Thus, the parameter matrices $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{Q}}$ is defined by the co-variances: a, b, c, d and e (Dropping the explicit dependence on current switch),

Table I Restoration results.

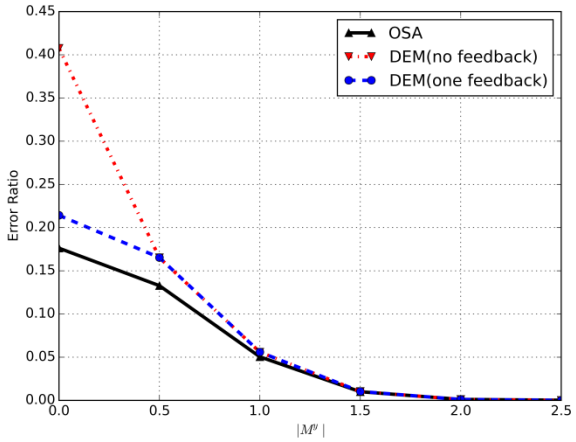
	Optimal Smoothing	Switching EM	Classical EM
MSE	0.158	0.175	0.341

Table II True and estimated parameters.

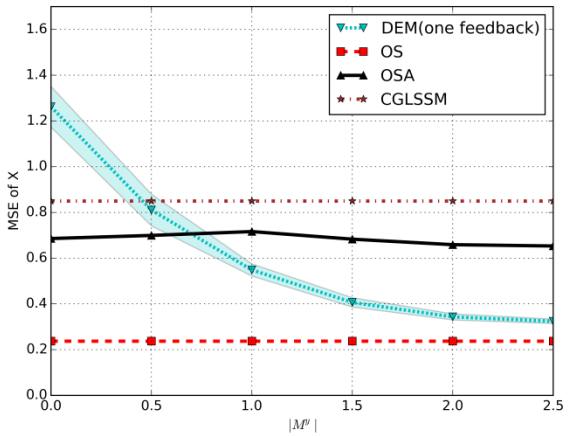
	Θ	$\Theta^{(0)}$	$\Theta^{(500)}_{SwitchingEM}$	$\Theta^{(500)}_{ClassicalEM}$
$\boldsymbol{\mathcal{F}}(R=0)$	$\begin{bmatrix}0.5 & 0.5\\ 1.0 & 0.0\end{bmatrix}$	$\begin{bmatrix}1.0 & 0.0\\ 1.0 & 0.0\end{bmatrix}$	$\begin{bmatrix}0.663 & 0.362\\ 1.077 & -0.164\end{bmatrix}$	$\begin{bmatrix}0.541 & 0.541\\ 0.647 & -0.132\end{bmatrix}$
$\boldsymbol{\mathcal{F}}(R=1)$	$\begin{bmatrix}0.2 & 0.8\\ 0.5 & 0.0\end{bmatrix}$		$\begin{bmatrix}0.261 & 0.721\\ 0.533 & -0.071\end{bmatrix}$	
$\boldsymbol{\mathcal{Q}}(R=0)$	$\begin{bmatrix}0.1 & 0.0\\ 0.0 & 0.5\end{bmatrix}$	$\begin{bmatrix}0.5 & 0.0\\ 0.0 & \mathrm{var}[\mathbf{Y}(\mathbf{R})]\end{bmatrix}$	$\begin{bmatrix}0.107 & 0.057\\ 0.057 & 0.475\end{bmatrix}$	$\begin{bmatrix}0.276 & 0.074\\ 0.074 & 0.317\end{bmatrix}$
$\boldsymbol{\mathcal{Q}}(R=1)$	$\begin{bmatrix}0.5 & 0.0\\ 0.0 & 0.1\end{bmatrix}$		$\begin{bmatrix}0.463 & 0.010\\ 0.010 & 0.093\end{bmatrix}$	

Table III $\mathbf{M}^y(\mathbf{R})$ estimated from DEM-CGPMSM.

$\lvert\mathbf{M}^y\rvert$	0.0	0.5	1.0	1.5	2.0	2.5
$(R=0)$	−0.004	0.539	1.014	1.505	2.001	2.501
$(R=1)$	−0.001	−0.405	−1.020	−1.518	−2.013	−2.512



(a) Error ratio of estimated R.



(b) MSE of restored X.

Fig. 2. Restoration results of series 2.

and when $d = cb$, we have $\boldsymbol{\mathcal{F}}^{yx} = (d - cb)/(1 - b^2) = 0$, the CGPMSM becomes a CGOMSM.

In this experiment, we set the co-variances to be: $b = 0.3$, $a^{l,0} = 0.1$, $a^{l,1} = 0.5$, $c^{l,0} = 0.4$, $c^{l,1} = 0.9$,

$e^{l,0} = 0.75,\ e^{l,1} = 0.33,\ d^{l,0}$ and $d^{l,1}$ are got by setting $\boldsymbol{\mathcal{F}}^{yx} = 0.4$. N = 2000 samples are generated with $|\mathbf{M}^y|$ ranging from 0.0 to 2.5, while $|\mathbf{M}^y|$ means the absolute value of the mean of $\mathbf{Y}$ corresponds to the two switches. For example: $|\mathbf{M}^y| = 2.5$ indicates that $\mathbf{M}^y(0) = 2.5$ and $\mathbf{M}^y(1) = -2.5$. Results are averages from 100 independent experiments, with 100 iterations for EM (PMC), 500 iterations for switching EM and valid feedback situation (the estimated feedback $\hat{\boldsymbol{\Gamma}}_y^{j,k}$ is positive semidefinite).

The error ratio of the estimated $\mathbf{R}_1^N$ through OSA and DEM-CGPMSM are shown in Fig. 2 (a). Comparing the error ratio result of DEM (no feedback) and DEM (one feedback), feedback from the Switching EM part effectively offers the proper initialization when $|\mathbf{M}^y|$ is near zero, which means a difficult situation for K-means to initialize the switch. In fact, EM is sensitive to the initialization and one feedback is enough for finding the proper one under this model. Fig. 2 (b) shows the restoration MSE of all methods applied. The confidence interval of DEM (one feedback) is displayed in light cyan. We can conclude that, when $|\mathbf{M}^y|$ increases, the switches are better estimated, and the DEM-CGPMSM gets better and steadier performance that can even surpass the OSA and CGLSSM, which are exact computation for approximation models. In fact, DEM-CGPMSM does have the chance to perform better than OSA and CGLSSM since it assumes no permanent 0 in parameters at special position of the model.

Table III reports the estimated value of $\mathbf{M}^y(\mathbf{R})$ with the value of $|\mathbf{M}^y|$ set for data simulation in first line.

4. CONCLUSION

An unsupervised parameter estimation and restoration method is proposed for "conditionally Gaussian pairwise Markov switching models" (CGPMSMs), which is based on two EM algorithms applied successively, named "double expectation-maximization-CGPMSM" (DEM-CGPMSM). The newly proposed partial algorithm (Switching EM) and the entire DEM-CGPMSM have been evaluated by simulated data. Results show that, the Switching EM, as an extension of the EM algorithm for switching models with known switches, can estimate the parameters effectively. The performance of DEM-CGPMSM can even surpass some sub-optimal supervised restoration methods.

Our future prospect includes: (i) the improvements on the initialization of the Switching EM part; (ii) the introduction of non-Gaussian distributions, taking into account recent results on hidden Markov chain and copula theory [15]; and (iii) the application of this method to non-linear real data.

5. REFERENCES

[1] N. Abbassi, D. Benboudjema, S. Derrode and W. Pieczynski, Optimal Filter Approximations in Conditionally Gaussian Pairwise Markov Switching Models, IEEE Trans. on Automatic Control, 60(4): 1104-1109, 2015.

[2] O. Cappé, E. Moulines, and T. Rydén, Inference in Hidden Markov Models, Springer-Verlag, New York, 2009.

[3] S. Derrode and W. Pieczynski, Exact Fast Computation of Optimal Filter in Gaussian Switching Linear Systems, IEEE Signal Processing Letters, 20(7): 701-704, 2013.

[4] W. Pieczynski and F. Desbouvries, Kalman Filtering using Pairwise Gaussian Models, Proceeding of the IEEE conference ICASSP, vol: 6, pp. 57-60, Hong-Kong, April 2003.

[5] B. Ait-el-Fquih, F. Desbouvries, Bayesian Smoothing Algorithms in Pairwise and Triplet Markov Chains, IEEE 13th Workshop on Statistical Signal Processing, pp. 721-726, Bordeau, France, July 2005.

[6] B. Ait-el-Fquih and F. Desbouvries, Unsupervised Signal Restoration in Partially Observed Markov Chains, Proceeding of the IEEE conference ICASSP, vol: 3, pp. III-III, May 2006.

[7] G. McLachlan, T. Krishnan, The EM Algorithm and Extensions. John Wiley & Sons, 2007.

[8] V. Némesin, S. Derrode, Robust Blind Pairwise Kalman Algorithms using QR Decompositions, IEEE Trans. on Signal Processing, 61(1): 5-9, 2013.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin, Maximum Likelihood from Incomplete Data via the EM Algorithm, Journal of the Royal Statistical Society, Series B (methodological), pp. 1-38, 1977.

[10] A. Doucet, N.J. Gordon and V. Krishnamurthy, Particle Filters for State Etimation of Jump Markov Linear Systems. IEEE Trans. on Signal Processing, 49(3): 613-24, 2001.

[11] B. Ristic, S. Arulampalam, and N. Gordon. Beyond the Kalman Filter: Particle Filters for Tracking Applications, Artech House, 2004.

[12] I. Gorynin, S. Derrode, E. Monfrini and W. Pieczynski, Exact Fast Smoothing in Switching Models with Application to Stochastic Volatility, Eusipco conference, Nice, France, August-September 2015.

[13] S. Derrode and W. Pieczynski, Signal and Image Segmentation using Pairwise Markov Chains, IEEE Trans. on Signal Processing, 52(9): 2477-89, 2004.

[14] L.E. Baum, T. Petrie, G. Soules, et al. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains, The Annals of Mathematical Statistics, 41(1): 164-171, 1970.

[15] S. Derrode and W. Pieczynski, Unsupervised Classification using Hidden Markov Chain with Unknown Noise Copulas and Margins, Signal Processing, Vol. 128, pp. 8-17, 2016.