Hidden-Markov models for time series of continuous proportions with excess zeros

Julien Alerini, Marie Cottrell, Madalina Olteanu

Hidden-Markov models for time series of continuous proportions with excess zeros

Julien Alerini[1], Marie Cottrell[2], and Madalina Olteanu[2]

[1] IHMC-PIREH, UMR 8589, Université Paris 1 Panthéon-Sorbonne, France
julien.alerini@univ-paris1.fr
[2] SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne, France
marie.cottrell, madalina.olteanu@univ-paris1.fr

Abstract. Bounded time series and time series of continuous proportions are often encountered in statistical modeling. Usually, they are addressed either by a logistic transformation of the data, or by specific probability distributions, such as Beta distribution. Nevertheless, these approaches may become quite tricky when the data show an over-dispersion in 0 and/or 1. In these cases, the zero-and/or-one Beta-inflated distributions, $\mathcal{ZOIB}$, are preferred. This manuscript combines $\mathcal{ZOIB}$ distributions with hidden-Markov models and proposes a flexible model, able to capture several regimes controlling the behavior of a time series of continuous proportions. For illustrating the practical interest of the proposed model, several examples on simulated data are given, as well as a case study on historical data, involving the military logistics of the Duchy of Savoy during the XVIth and the XVIIth centuries.

1 Introduction

Time series of continuous proportions or percentages are often encountered in various research fields such as economy, biology or history. For instance, one may be interested in modeling the fraction of income a family devotes to lodging or taxes; or in modeling the proportion of a population exposed to fine particles pollution or subject to a certain type of disease. In our case, as it will be shown later, we are interested in understanding the rhythms of the Sabaudian State during the XVIth and the XVIIth centuries, and more particularly the evolution of the ratio of legislative texts issued by the Duchy and related to military logistics, among the entire production of law.

In statistical modeling, the two common approaches for dealing with continuous proportions are, on the one hand, a logistic transformation of the data [1], and, on the other hand, the use of specific probability distributions such as Beta or Dirichlet, [2]. However, both of these approaches have a major drawback, since they do not take into account the possibility of an over-dispersion in the limit values, 0 and/or 1. During the last few years, this issue has been addressed by several authors, who proposed either further transforming the data [3], or introducing specific probability masses in 0 and/or 1, hence using zero-and/or-one Beta

Inflated distributions. The latter approach has been intensively studied during the last five years, mainly in a regression context [4], [5].
In this manuscript, we aim at using the Beta inflated distributions in a framework different from that of regression, since our main interest is to uncover and to highlight the possible existence of several regimes in a time series of proportions. With this in mind, we introduce a hidden-Markov model, having as emission distribution a Beta inflated distribution, $\mathcal{ZOIB}$-HMM in abbreviated form. Originally introduced for speech recognition [6], hidden Markov models (HMM hereafter) are especially interesting in the context of the presumed existence of several regimes controlling the parameters of the model, or the parameters of the emission distribution.
The next sections are organized as follows: in Section 2, we recall the definition and the properties of the zero-and-one Beta-inflated distribution and then, in Section 3, we introduce the $\mathcal{ZOIB}$-HMM model and describe the estimation procedure, which is essentially an expectation-maximisation (EM) scheme. Section 4 contains several experimental results, with a discussion on the convergence properties, the speed of the algorithm and the possible identifiability issues, while Section 5 is devoted to presenting the results on a real data set, coming from medieval history. Finally, a conclusion will follow in Section 6.

2 Zero-and-one Beta-inflated distributions

As mentioned in the introduction, statistical models based on Beta distributions assume the data to be valued in the open interval $]0, 1[$. In practical applications, this is rarely the case, and the situation of an over-dispersion in 0 and/or 1 appears quite often. The solution for dealing with this is to mix the Beta distribution either with a Dirac mass (in 0 for data valued in $[0, 1[$, in 1 for data valued in $]0, 1]$), or with a Bernoulli distribution (for data valued in $[0, 1]$). Only the latter case will be addressed here, but the reader may refer to [7] for a complete review of Beta inflated distributions.
The probability density function of a Beta distribution with parameters $\alpha, \beta > 0$, denoted $\mathcal{B}e(\alpha, \beta)$, is

$$f_B\left(x; \alpha, \beta\right) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}_{x\in]0,1[}, \tag{1}$$

where Γ is the Gamma function. Also, if $X \sim \mathcal{B}(\alpha, \beta)$ and $\mu = \frac{\alpha}{\alpha+\beta}$, $\phi = \alpha + \beta$, the expectation and the variance of X may be expressed as :

$$\mathbb{E}(X) = \mu \ ; \ \mathbb{V}(X) = \frac{\mu(1-\mu)}{\phi + 1}. \tag{2}$$

Besides its support reduced to the interval $]0, 1[$, the interest of using a Beta distribution for statistical modeling also resides in the large variety of shapes for its density, which makes it quite appealing for applications (see Figure 1).
Let us now define the zero-and-one Beta-inflated as a mixture between a Bernoulli and a Beta distribution, using a latent variable Y. If $Y \sim \mathcal{B}(\eta)$,

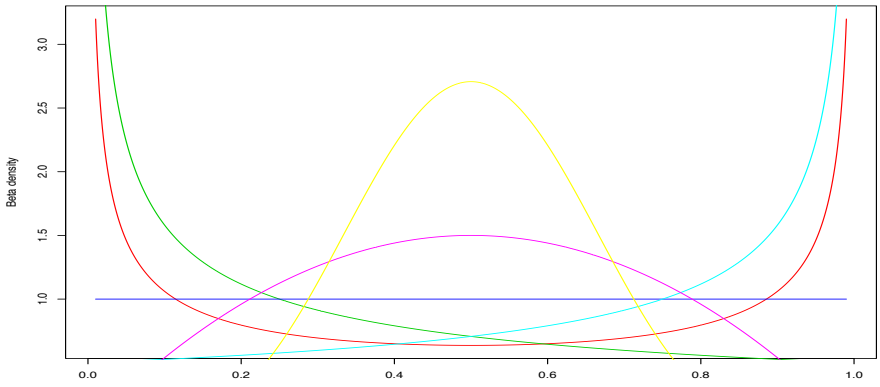

Fig. 1. Examples of Beta distributions for various parameters α and β.

define X conditionally to Y such that $X|(Y=1) \sim \mathcal{B}(\gamma)$ and $X|(Y=0) \sim \mathcal{B}e(\alpha,\beta)$. The marginal density[3] of X is

$$f(x;\eta,\gamma,\alpha,\beta) = (\eta\gamma)^{\mathbb{1}_{x=1}}(\eta(1-\gamma))^{\mathbb{1}_{x=0}}((1-\eta)f_B(x;\alpha,\beta))^{\mathbb{1}_{x\in]0,1[}}, \quad (3)$$

where $\eta \in]0,1[$ is the mixture parameter, $\gamma \in]0,1[$ is the Bernoulli-distribution parameter and $\alpha,\beta > 0$ are the Beta-distribution parameters. Throughout the rest of the paper, let $\xi = (\eta,\gamma,\alpha,\beta)$ be the four-dimensional parameter of a zero-and-one Beta-inflated distribution, $\mathcal{ZOIB}(\xi)$.
Consider now $X_1^T = (X_1,...,X_T)$ an i.i.d. T-sample of $\mathcal{ZOIB}(\xi)$. The likelihood may be written as :

$$\mathcal{L}(X_1^T;\xi) = \prod_{t=1}^{T}\left[\eta^{\mathbb{1}_{X_t\in\{0,1\}}}(1-\eta)^{\mathbb{1}_{X_t\in]0,1[}}\right] \times \prod_{t=1}^{T}\left[\gamma^{\mathbb{1}_{X_t=1}}(1-\gamma)^{\mathbb{1}_{X_t=0}}\right]$$
$$\times \prod_{t=1}^{T} f_B(X_t;\alpha,\beta)^{\mathbb{1}_{X_t\in]0,1[}} = \mathcal{L}_1(X_1^T;\eta)\mathcal{L}_2(X_1^T;\gamma)\mathcal{L}_3(X_1^T;\alpha,\beta) \quad (4)$$

Maximizing the likelihood consists in maximizing each of the three terms in the product, which are independent in the parameter components. For the mixture parameter η, the maximum likelihood estimate (MLE) is computed by maximizing

$$\ln \mathcal{L}_1(X_1^T;\eta) = \ln\eta \sum_{t=1}^{T}\mathbb{1}_{X_t\in\{0,1\}} + \ln(1-\eta)\sum_{t=1}^{T}\mathbb{1}_{X_t\in]0,1[}, \quad (5)$$

[3] the density is taken with respect to the probability measure $\lambda + \delta_0 + \delta_1$, where λ is the Lebesgue measure on $[0,1]$, and δ_0 and δ_1 are Dirac masses in 0 and 1.

which yields

$$\hat{\eta} = \frac{\sum_{t=1}^{T} \mathbb{1}_{X_t \in \{0,1\}}}{\sum_{t=1}^{T} \mathbb{1}_{X_t \in \{0,1\}} + \sum_{t=1}^{T} \mathbb{1}_{X_t \in]0,1[}} = \frac{T_1}{T}, \tag{6}$$

where $T_1 = \sum_{t=1}^{T} \mathbb{1}_{X_t \in \{0,1\}}$. For the Bernoulli parameter γ, the MLE is computed by maximizing

$$\ln \mathcal{L}_2(X_1^T; \gamma) = \ln \gamma \sum_{t=1}^{T} \mathbb{1}_{X_t=1} + \ln(1-\gamma) \sum_{t=1}^{T} \mathbb{1}_{X_t=0}, \tag{7}$$

which yields

$$\hat{\gamma} = \frac{\sum_{t=1}^{T} \mathbb{1}_{X_t=1}}{\sum_{t=1}^{T} \mathbb{1}_{X_t=1} + \sum_{t=1}^{T} \mathbb{1}_{X_t=0}} = \frac{T_2}{T_1}, \tag{8}$$

where $T_2 = \sum_{t=1}^{T} \mathbb{1}_{X_t=1}$. Finally, for the Beta parameters, α and β, one has to maximize

$$\ln \mathcal{L}_3(X_1^T; \alpha, \beta) = \sum_{X_t \in]0,1[} \ln f_B(X_t; \alpha, \beta). \tag{9}$$

In this case, since there is no analytical form for the MLE of a Beta distribution, the solution may be found using numerical optimization. However, in order to avoid numerical issues linked to the initial values of the gradient-descent based algorithms, an approximation of the MLE with the moment estimates is preferred. Following Equation 2, the moment estimates of α and β are :

$$\tilde{\alpha} = \tilde{\mu}\tilde{\phi} \ , \ \tilde{\beta} = (1-\tilde{\mu})\tilde{\phi}, \tag{10}$$

where

$$\tilde{\mu} = \frac{1}{n - T_1} \sum_{X_t \in]0,1[} X_t \ , \ \tilde{\phi} = \frac{\tilde{\mu}(1-\tilde{\mu})}{s^2} - 1,$$
$$\text{and } s^2 = \frac{1}{n - T_1} \sum_{X_t \in]0,1[} (X_t - \tilde{\mu})^2 \, . \tag{11}$$

3 Estimation procedure for the $\mathcal{ZOIB}$-HMM model

In this section, the parameters of a zero-and-one Beta-inflated distribution are supposed to be controlled by a hidden Markov chain with a finite number of states.

3.1 The model

Let $(X_t)_{t\in\mathbb{Z}}$ be the observed time series, valued in $[0,1]$, and let $(S_t)_{t\in\mathbb{N}}$ be the unobserved process, controlling the parameters of the distribution of X_t. Throughout the rest of the paper, S_t is supposed to be a homogeneous Markov chain, irreducible, recurrent and aperiodic, valued in a finite state-space $E = \{e_1, ..., e_q\}$ and defined by its transition matrix $\Pi = (\pi_{ij})_{i,j=1,...,q}$,

$$\pi_{ij} = \mathbb{P}\left(S_t = e_j | S_{t-1} = e_i\right),$$

with $\pi_{ij} > 0$, $\sum_{j=1}^{q} \pi_{ij} = 1$, and by its initial probability distribution π^0, $\pi_i^0 = \mathbb{P}\left(S_1 = e_i\right)$, $\forall i = 1, ..., q$.
Furthermore, let us suppose that X_t are independent conditionally to S_t, and that X_t conditionally to S_t is distributed according to a zero-and-one Beta-inflated distribution, $\mathcal{ZOIB}(\xi_i)$, with $\xi_i = (\eta_i, \gamma_i, \alpha_i, \beta_i) \in]0,1[^2 \times]0, +\infty[^2$. For a fixed number of states q in the hidden Markov chain, the set of possible values for the parameters may be written as:

$$\Theta = \Big\{ \theta = ((\xi_i)_{i=1,...,q}, \Pi) \in (]0,1[^2 \times]0,+\infty[^2)^q \times]0,1[^{q^2}, \\ \text{and } \forall i \in \{1 \cdots q\}, \sum_{j=1}^{q} \pi_{ij} = 1 \Big\}. \quad (12)$$

3.2 The EM algorithm for $\mathcal{ZOIB}$-HMM

Since the above model involves a hidden Markov chain, the estimation procedure is carried out using the EM algorithm [6], [8]. With the previous assumptions, and with the notations $X_1^T = (X_1, ..., X_T)$, $S_1^T = (S_1, ..., S_T)$, the complete likelihood is given by:

$$\mathcal{L}(X_1^T, S_1^T; \theta) = \prod_{t=1}^{T} \prod_{i=1}^{q} f(X_t | S_t = e_i; \xi_i)^{\mathbb{1}_{\{S_t = e_i\}}} \prod_{t=2}^{T} \prod_{i,j=1}^{q} \pi_{ij}^{\mathbb{1}_{\{S_{t-1}=e_i, S_t=e_j\}}} \times C, \quad (13)$$

where $f(X_t | S_t = e_i; \xi_i)$ is the Beta-inflated density, conditionally to the hidden state $S_t = e_i$ and defined in Equation 3, and $C = \prod_{i=1}^{q} \left(\pi_i^0\right)^{\mathbb{1}_{\{S_t=e_i\}}}$ is the likelihood of the initial state of the Markov chain.
When expressing the Beta-inflated density in its analytical form, the complete likelihood may further be written as:

$$\mathcal{L}(X_1^T, S_1^T; \theta) = \prod_{t=1}^{T} \prod_{i=1}^{q} \left(\eta_i^{\mathbb{1}_{X_t \in \{0,1\}}} (1-\eta_i)^{\mathbb{1}_{X_t \in]0,1[}} \right)^{\mathbb{1}_{\{S_t=e_i\}}} \times \\ \prod_{t=1}^{T} \prod_{i=1}^{q} \left(\gamma_i^{\mathbb{1}_{X_t=1}} (1-\gamma_i)^{\mathbb{1}_{X_t=0}} \right)^{\mathbb{1}_{\{S_t=e_i\}}} \times \\ \prod_{t=1}^{T} \prod_{i=1}^{q} \left(f_B\left(X_t, \alpha_i, \beta_i\right) \right)^{\mathbb{1}_{\{X_t \in]0,1[, S_t=e_i\}}} \times \prod_{t=2}^{T} \prod_{i,j=1}^{q} \pi_{ij}^{\mathbb{1}_{\{S_{t-1}=e_i, S_t=e_j\}}} \times C \\ = \mathcal{L}_1(X_1^T, S_1^T; \boldsymbol{\eta}) \mathcal{L}_2(X_1^T, S_1^T; \boldsymbol{\gamma}) \mathcal{L}_3(X_1^T, S_1^T; \boldsymbol{\alpha}, \boldsymbol{\beta}) \mathcal{L}_4(X_1^T, S_1^T; \Pi) \quad (14)$$

where $\boldsymbol{\eta} = (\eta_1, ..., \eta_q)$, $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_q)$, $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_q)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_q)$. The EM algorithm consists in iteratively maximizing the expected value of the complete likelihood with respect to θ and conditionally to the observed data set X_1^T and a fixed value of the parameter $\theta^\star$, and then updating $\theta^\star$ at each step.

E-Step. The E-step is given by the computation of the expected value of the complete likelihood, conditionally to the observed data,

$$Q(\theta|\theta^\star) = \mathbb{E}_{\theta^\star}\left[\ln \mathcal{L}(X_1^T, S_1^T; \theta)|X_1^T\right]. \tag{15}$$

According to Equation 14, $Q(\theta|\theta^\star)$ can be split into :

$$Q(\theta|\theta^\star) = Q_1(\boldsymbol{\eta}|\theta^\star) + Q_2(\boldsymbol{\gamma}|\theta^\star) + Q_3(\boldsymbol{\alpha}, \boldsymbol{\beta}|\theta^\star) + Q_4(\Pi|\theta^\star)\ , \tag{16}$$

where

$$\begin{aligned} Q_1(\boldsymbol{\eta}|\theta^\star) &= \mathbb{E}_{\theta^\star}\left[\ln \mathcal{L}_1(X_1^T, S_1^T; \boldsymbol{\eta})|X_1^T\right] \\ &= \sum_{i=1}^{q}\left[\sum_{X_t \in \{0,1\}} \omega_t(e_i) \ln \eta_i + \sum_{X_t \in]0,1[} \omega_t(e_i) \ln(1-\eta_i)\right], \end{aligned} \tag{17}$$

with $\omega_t(e_i) = \mathbb{P}_{\theta^\star}(S_t = e_i|X_1^T)$;

$$\begin{aligned} Q_2(\boldsymbol{\gamma}|\theta^\star) &= \mathbb{E}_{\theta^\star}\left[\ln \mathcal{L}_2(X_1^T, S_1^T; \boldsymbol{\gamma})|X_1^T\right] \\ &= \sum_{i=1}^{q}\left[\sum_{X_t=1} \omega_t(e_i) \ln \gamma_i + \sum_{X_t=0} \omega_t(e_i) \ln(1-\gamma_i)\right]; \end{aligned} \tag{18}$$

$$\begin{aligned} Q_3(\boldsymbol{\alpha}, \boldsymbol{\beta}|\theta^\star) &= \mathbb{E}_{\theta^\star}\left[\ln \mathcal{L}_3(X_1^T, S_1^T; \boldsymbol{\alpha}, \boldsymbol{\beta})|X_1^T\right] \\ &= \sum_{i=1}^{q} \sum_{X_t \in]0,1[} \omega_t(e_i) \ln\left(f_B\left(X_t, \alpha_i, \beta_i\right)\right); \end{aligned} \tag{19}$$

$$Q_4(\Pi|\theta^\star) = \mathbb{E}_{\theta^\star}\left[\ln \mathcal{L}_4(X_1^T, S_1^T; \Pi)|X_1^T\right] = \sum_{i,j=1}^{q} \sum_{t=2}^{T} \omega_t(e_i, e_j) \ln \pi_{ij}\ , \tag{20}$$

with $\omega_t(e_i, e_j) = \mathbb{P}_{\theta^\star}(S_{t-1} = e_i, S_t = e_j|X_1^T)$. The probabilities $\omega_t(e_i)$ and $\omega_t(e_i, e_j)$ may be easily computed using the forward-backward procedure, typical for the EM algorithm [9].

M-Step. Thanks to the factorization of the complete likelihood, the optimization step may be performed by independently maximizing each term of $Q(\theta|\theta^\star)$. For η_i, γ_i and π_{ij}, the following analytical expressions are straightforward:

$$\hat{\eta}_i = \frac{\sum_{X_t \in \{0,1\}} \omega_t(e_i)}{\sum_{t=1}^{T} \omega_t(e_i)}\ ,\ \hat{\gamma}_i = \frac{\sum_{X_t=1} \omega_t(e_i)}{\sum_{X_t \in \{0,1\}} \omega_t(e_i)}\ ,$$
$$\text{and } \hat{\pi}_{ij} = \frac{\sum_{t=2}^{T} \omega_t(e_i, e_j)}{\sum_{t=1}^{T} \omega_t(e_i)}\ . \tag{21}$$

As for α_i and β_i, there are no analytical expressions of the estimates, directly tractable from $Q_3(\boldsymbol{\alpha}, \boldsymbol{\beta}|\theta^\star)$. Rather than numerically optimizing this function, operation which would slow down the algorithm and possibly introduce numerical instability, we prefer the use of moment estimates, which appear as good substitues for the MLE:

$$\tilde{\alpha}_i = \tilde{\mu}_i\tilde{\phi}_i \ , \ \tilde{\beta}_i = (1-\tilde{\mu}_i)\tilde{\phi}_i, \tag{22}$$

where

$$\tilde{\mu}_i = \frac{\sum_{X_t\in]0,1[}\omega_t(e_i)X_t}{\sum_{X_t\in]0,1[}\omega_t(e_i)} \ , \ \tilde{\phi}_i = \frac{\tilde{\mu}_i(1-\tilde{\mu}_i)}{s_i^2} - 1,$$
$$s_i^2 = \frac{\sum_{X_t\in]0,1[}\omega_t(e_i)\left(X_t-\tilde{\mu}_i\right)^2}{\sum_{X_t\in]0,1[}\omega_t(e_i)}. \tag{23}$$

As one may easily notice, Equations 22 and 23 are very similar to Equations 10 and 11, except for the weights $\omega_t(e_i)$, which are introduced by the hidden Markov chain and represent the conditional probabilities of being in state e_i at time t, given the observed data, X_1^T.

4 Experimental results

In order to test the quality of the estimates and its convergence rate, the algorithm was trained on several simulated examples. For each of the following scenarios and for sample sizes ranging from 500 to 1 000, 100 different trajectories of a two-state ($q = 2$) $\mathcal{ZOIB}$-HMM are simulated. The values of the parameters used for the simulations are the following :

$$\mathit{\Pi} = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}, (\alpha_1, \alpha_2) = (1; 0.5), (\beta_1, \beta_2) = (1; 2), (\gamma_1, \gamma_2) = (0.5; 0.9),$$

and $(\eta_1, \eta_2) = (\eta_1, 0.8)$, where $\eta_1 \in \{0.1, 0.3, 0.5, 0.7\}$. The results are detailed in Tables 1, 2, 3 and 4 below. In each case are reported the mean values of the estimates, as well as their standard errors and medians. We also provide the squared bias and the ratio of errors in the a posteriori identification of the hidden regimes (mean-values, standard errors and medians).

According to these first results on synthetic data, most of the parameters ($\mathit{\Pi}$, the η's and the γ's) are generally correctly estimated, even for short time series. However, the quality of the estimated transition matrix diminishes when η_1 has larger values (the ratio of zeros and ones is overriding the ratio of values in $]0, 1[$). The α's and the β's are correctly estimated for small values of η_1 and sufficiently large time series, with a length at least equal to 1 000. However, the algorithm fails in fairly approaching them when η_1 is greater than 0.5 or for time series shorter than 1 000 observations.

Furthermore, when comparing the mean values of the estimates with their medians, one may easily see that, if considering the medians, the performances of the algorithm are eventually not bad even in this limit cases. When looking into details, some of the incoherences come from

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.82(*0.18*)	0.18(*0.18*)	0.84(*0.16*)	0.16(*0.16*)
	0.89	**0.11**	**0.89**	**0.11**
	0.19(*0.20*)	0.81(*0.20*)	0.16(*0.16*)	0.84(*0.16*)
	0.11	**0.89**	**0.11**	**0.89**
$\hat{\alpha}_1, \hat{\alpha}_2$	0.99(*0.27*)	0.74(*0.66*)	0.99(*0.14*)	0.54(*0.15*)
	0.96	**0.60**	**1.00**	**0.52**
$\hat{\beta}_1, \hat{\beta}_2$	0.98(*0.11*)	2.70(*1.83*)	1.00(*0.11*)	2.11(*0.75*)
	0.99	**2.35**	**1.00**	**2.09**
$\hat{\gamma}_1, \hat{\gamma}_2$	0.55(*0.22*)	0.88(*0.09*)	0.52(*0.18*)	0.89(*0.07*)
	0.53	**0.90**	**0.51**	**0.90**
$\hat{\eta}_1, \hat{\eta}_2$	0.15(*0.13*)	0.76(*0.15*)	0.15(*0.14*)	0.77(*0.13*)
	0.10	**0.80**	**0.10**	**0.80**
$Bias(\theta)^2$	1.39(*1.65*)		0.71(*0.57*)	
	0.89		**0.53**	
$\%ERR$	14.6(*18.4*)		12.1(*14.8*)	
	6.9		**6.7**	

Table 1. Simulation results for $\eta_1 = 0.1$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.78(*0.19*)	0.22(*0.19*)	0.81(*0.17*)	0.19(*0.17*)
	0.86	**0.14**	**?**	**?**
	0.24(*0.24*)	0.76(*0.24*)	0.18(*0.17*)	0.82(*0.17*)
	0.11	**0.89**	**?**	**?**
$\hat{\alpha}_1, \hat{\alpha}_2$	1.06(*0.78*)	1.12(*3.49*)	0.96(*0.14*)	0.56(*0.15*)
	0.97	**0.56**	**?**	**?**
$\hat{\beta}_1, \hat{\beta}_2$	1.01(*0.18*)	2.91(*3.30*)	0.98(*0.11*)	1.99(*0.77*)
	0.99	**1.88**	**?**	**?**
$\hat{\gamma}_1, \hat{\gamma}_2$	0.51(*0.21*)	0.86(*0.15*)	0.54(*0.19*)	0.86(*0.10*)
	0.48	**0.89**	**?**	**?**
$\hat{\eta}_1, \hat{\eta}_2$	0.34(*0.14*)	0.75(*0.16*)	0.32(*0.14*)	0.77(*0.13*)
	0.29	**0.80**	**?**	**?**
$Bias(\theta)^2$	2.23(*4.51*)		0.76(*0.55*)	
	0.94		**?**	
$\%ERR$	21.2(*16.5*)		18.2(*16.0*)	
	12.2		**?**	

Table 2. Simulation results for $\eta_1 = 0.3$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

atypical time series in the simulations which rise identifiability issues. In all cases, the experimental section should be further improved with more examples, involving more diverse scenarios for the parameters and longer time series.

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.69(*0.23*)	0.31(*0.23*)	0.77(*0.19*)	0.23(*0.19*)
	0.76	**0.24**	**0.88**	**0.12**
	0.24(*0.19*)	0.76(*0.19*)	0.24(*0.22*)	0.76(*0.22*)
	0.18	**0.82**	**0.13**	**0.87**
$\hat{\alpha}_1, \hat{\alpha}_2$	13.08(*80.91*)	4.01(*27.24*)	0.94(*0.19*)	0.58(*0.30*)
	0.93	**0.60**	**0.97**	**0.54**
$\hat{\beta}_1, \hat{\beta}_2$	4.30(*20.91*)	7.58(*45.68*)	0.98(*0.12*)	2.27(*2.59*)
	1.00	**1.74**	**0.98**	**1.71**
$\hat{\gamma}_1, \hat{\gamma}_2$	0.50(*0.24*)	0.81(*0.20*)	0.54(*0.21*)	0.83(*0.16*)
	0.48	**0.88**	**0.51**	**0.89**
$\hat{\eta}_1, \hat{\eta}_2$	0.51(*0.22*)	0.75(*0.18*)	0.50(*0.13*)	0.76(*0.15*)
	0.49	**0.79**	**0.50**	**0.79**
$Bias(\theta)^2$	20.31(*98.05*)		1.25(*2.39*)	
	1.34		**0.79**	
$\%ERR$	30.6(*18.7*)		26.4(*16.2*)	
	23.2		**18.2**	

Table 3. Simulation results for $\eta_1 = 0.5$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

	$T = 500$		$T = 1000$	
$\hat{\Pi}$	0.59(*0.25*)	0.41(*0.25*)	0.61(*0.24*)	0.39(*0.24*)
	0.60	**0.40**	**0.61**	**0.39**
	0.33(*0.21*)	0.67(*0.21*)	0.30(*0.19*)	0.70(*0.19*)
	0.26	**0.74**	**0.27**	**0.73**
$\hat{\alpha}_1, \hat{\alpha}_2$	4.15(*31.39*)	10.58(*53.98*)	0.89(*0.37*)	5.96(*51.40*)
	0.81	**0.56**	**0.86**	**0.54**
$\hat{\beta}_1, \hat{\beta}_2$	2.15(*11.52*)	80.26(*664.83*)	0.97(*0.20*)	5.77(*36.67*)
	0.99	**1.53**	**0.99**	**1.56**
$\hat{\gamma}_1, \hat{\gamma}_2$	0.51(*0.26*)	0.80(*0.19*)	0.53(*0.24*)	0.81(*0.16*)
	0.53	**0.83**	**0.50**	**0.87**
$\hat{\eta}_1, \hat{\eta}_2$	0.64(*0.23*)	0.77(*0.19*)	0.64(*0.19*)	0.78(*0.17*)
	0.69	**0.80**	**0.68**	**0.80**
$Bias(\theta)^2$	85.93(*666.99*)		7.86(*63.01*)	
	1.46		**1.24**	
$\%ERR$	37.8(*15.2*)		30.0(*15.0*)	
	35.8		**31.0**	

Table 4. Simulation results for $\eta_1 = 0.7$ and 100 time-series of length T. Mean, standard error (italics) and median (bold) of the estimates.

5 A case study on historical data

The aim of this section is to apply the proposed model in studying the rhythms of the Duchy of Savoy, during the XVIth and the XVIIth centuries. These two centuries were marked by deep political changes and by several long and intense wars. It was a period during which the Duchy

changed and shaped its structure and its functioning as a state. More specifically, we are interested in the production of legislative texts related to military logistics, compared to the entire production of law. The corpus of data comes from the massive work of F-A. Duboin [10] [11]. According to [11], this edition would be exhaustive, and few texts would be missing. Between 1559 and 1661, the State as a whole issued 55.5 law texts per year (in average), of which 4.8 in connection with military logistics. The ratio between the texts on military logistics and the total number of texts of law varies between 0 and 0.23, and, as one may see in Figure 2, the importance of military logistics is far from negligible.

In order to evaluate the temporality of the State, the selected corpus of documents is thus represented as a time series. After having considered several representations (yearly, quarterly, monthly), the monthly approach was selected, as being sufficiently fine for one to observe the closeness between making the decision and issuing the associated text of law. A full description of the corpus and of its construction is available in [12]. In a previous work [13], the series of texts related to military logistics only was studied, using hidden Markov models for count data. The results appeared as very encouraging and they allowed to point out several specificities of the historical period of interest such as short-term events, but also a long transition between two normative systems of the Duchy. However, in [13] the relation between the texts on military logistics and the entire production of law was left out. In this manuscript, we focus on this ratio, computed between the number texts on military logistics and the complete production of law, recorded each month. The resulting series is of length 1 236 and has an important over-dispersion in 0.

After having trained a two-state hidden Markov model with Beta-inflated distributions, the estimated parameters are the following :

$$\hat{\Pi} = \begin{pmatrix} 0.83 & 0.17 \\ 0.26 & 0.74 \end{pmatrix} \ , \ (\hat{\alpha}_1, \hat{\alpha}_2) = (5.92; 4.11) \ , \ (\hat{\beta}_1, \hat{\beta}_2) = (8.07; 15.83),$$

$(\hat{\gamma}_1, \hat{\gamma}_2) = (0.01; 0.02)$ and $(\hat{\eta}_1, \hat{\eta}_2) = (0.87, 0.45)$.

The Viterbi algorithm allows to track the a posteriori probabilities of the hidden states and to represent the estimated trajectory of the hidden Markov chain, as shown in Figure 3. The results are globally consistent with our previous findings in [13]. The second regime is more persistent during the XVIIth century, which mainly corresponds to the end of the long transition period found in our previous work. However, the two regimes do not appear as stable and the Markov chain is changing often from one state to another. From this point of view, these results on the ratio series appear as less convincing than the results on logistic texts only. Some adjustments should thus be made in order to improve this and one way to tackle this would be to restrict the Beta-inflated distributions in the HMM either to 0 or to 1 additional probability masses. Indeed, this assumption would favor the hypothesis of a regime with an intense production of texts on military logistics versus a regime with a low production.

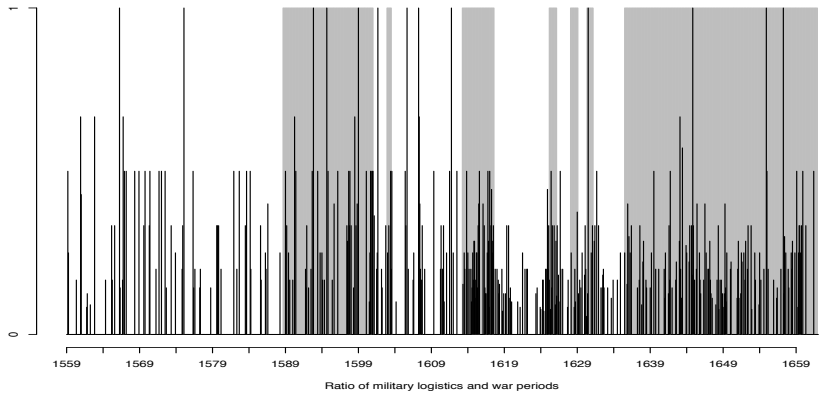

Fig. 2. Ratio of texts on military logistics among the entire production of law. The periods of war for the Duchy are in grey.

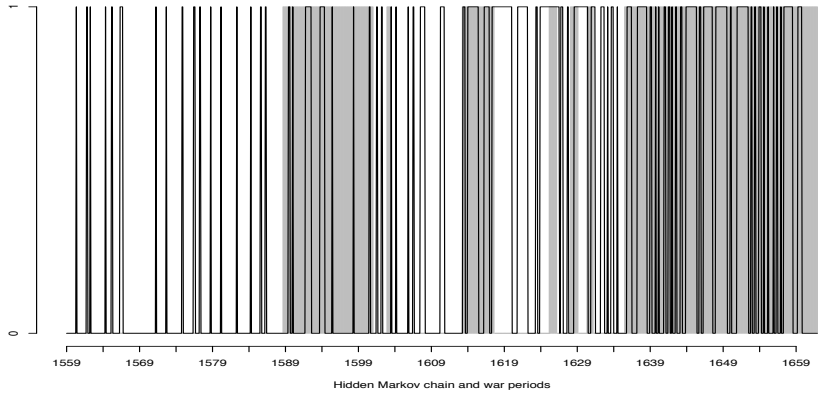

Fig. 3. A posteriori estimated hidden states of the model. The periods of war for the Duchy are in grey.

6 Conclusion

The present manuscript introduced the Beta-inflated distributions in the framework of hidden-Markov models. The results on simulated examples showed that the EM algorithm usually manages quite well in estimating the parameters of the model, provided the time series is sufficiently long

and provided the parameters of the regimes are sufficiently different. We aim at further improving this section, by adding more, and more various, examples. Finally, the results on the real dataset were consistent with previous findings, although less convincing in terms of stability of the regimes. We are currently studying the possibility of restricting the Beta-inflated distribution in the model to 0 or to 1 additional probability masses.

References

1. Wallis, K.F.: Time series analysis of bounded economic variables. Journal of Time Series Analysis **8**(1) (1987) 115–123
2. Ferrari, S., Cribari-Neto, F.: Beta regression for modelling rates and proportions. Journal of Applied Statistics **31**(7) (2004) 799–815
3. Smithson, M., Verkuilen, J.: A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. Psychological methods **11**(1) (2006) 54
4. Simas, A.B., Barreto-Souza, W., Rocha, A.V.: Improved estimators for a general class of beta regression models. Computational Statistics & Data Analysis **54**(2) (2010) 348 – 366
5. Ospina, R., Ferrari, S.L.: A general class of zero-or-one inflated beta regression models. Computational Statistics & Data Analysis **56**(6) (2012) 1609 – 1623
6. Baum, L., Petrie, T.: Statistical inference for probabilistic functions of finite state markov chains. The Annals of Mathematical Statistics **37**(6) (1966) 1554–1563
7. Ospina, R., Ferrari, S.L.P.: Inflated beta distributions. Statistical Papers **51**(1) (2008) 111
8. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society (B) **39**(1) (1977) 1–38
9. Rabiner, L.R.: A tutorial on hidden markov models and selected applications in speech recognition. Proceedings of the IEEE **77**(2) (Feb 1989) 257–286
10. Duboin, F.A.: Raccoolta per ordine di materie delle leggi cio editti, manifesti, ecc., pubblicati negli stati della Real Casa di Savoia fino all'8 dicembre 1798, Torino (1818-1869)
11. Couzin, T.: Contribution piémontaise à la genèse de l'État italien. L'historicité de la Raccolta per ordine di materie delle leggi (1818-1868). Bolettino Storico-Bibliografico Subalpino **CVI** (2008) 101–120
12. Alerini, J.: La Savoie et le "Chemin espagnol", les communautés alpines à l'épreuve de la logistique militaire (1560-1659). PhD thesis, Université Paris 1 Panthéon-Sorbonne (2012)
13. Alerini, J., Olteanu, M., Ridgway, J.: Markov and the Duchy of Savoy: segmenting a century with regime-switching models. working paper or preprint (January 2017)