DiscreteTS : two hidden-Markov models for time series of count data

J. Alerini[1], M. Olteanu[2], J. Ridgway[2]

[1]PIREH-LAMOP, Université Paris 1

[2]SAMM (Statistique, Analyse et Modélisation Multidisciplinaire), Université Paris 1

1ères Rencontres R
2-3 juillet 2012, Bordeaux

The data

Our sources :

- Duboin, Felice-Amato, *Raccolta per ordine di materie delle leggi cioè editti, manifesti, ecc., pubblicati negli stati della Real Casa di Savoia fino all'8 dicembre 1798*, Turin, 1818-1869, 30 volumes

- A dataset containing 472 texts about logistics among 5775 documents issued between 1559 and 1660 (8,17% of the legislation)

Theoretical issues in history

- Sources quality
- Processing the historical information

The data

Our sources :

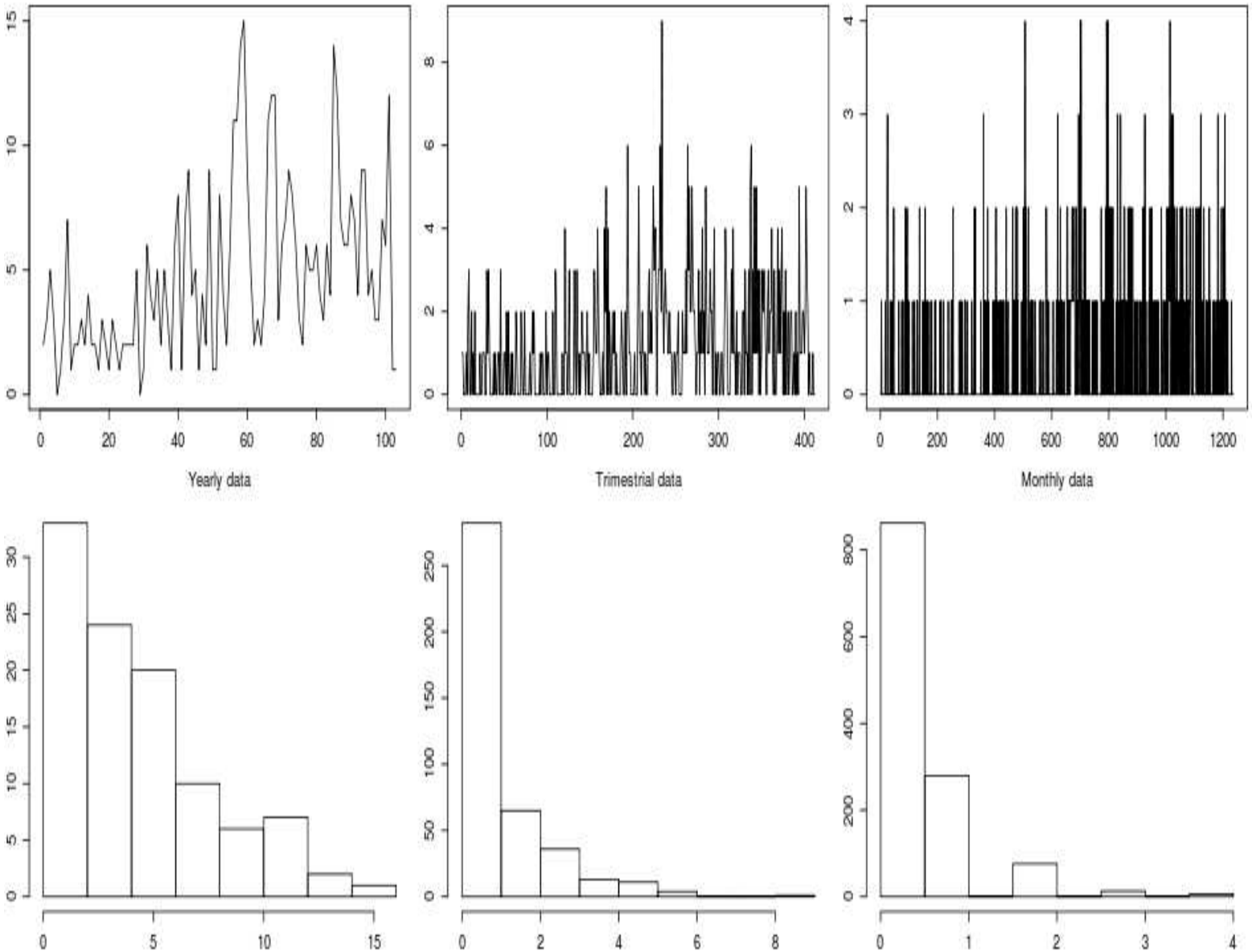
- Duboin, Felice-Amato, *Raccolta per ordine di materie delle leggi cioè editti, manifesti, ecc., pubblicati negli stati della Real Casa di Savoia fino all'8 dicembre 1798*, Turin, 1818-1869, 30 volumes

- A dataset containing 472 texts about logistics among 5775 documents issued between 1559 and 1660 (8,17% of the legislation)

Theoretical issues in history

- Sources quality
- Processing the historical information

Which time-scale for the analysis ?

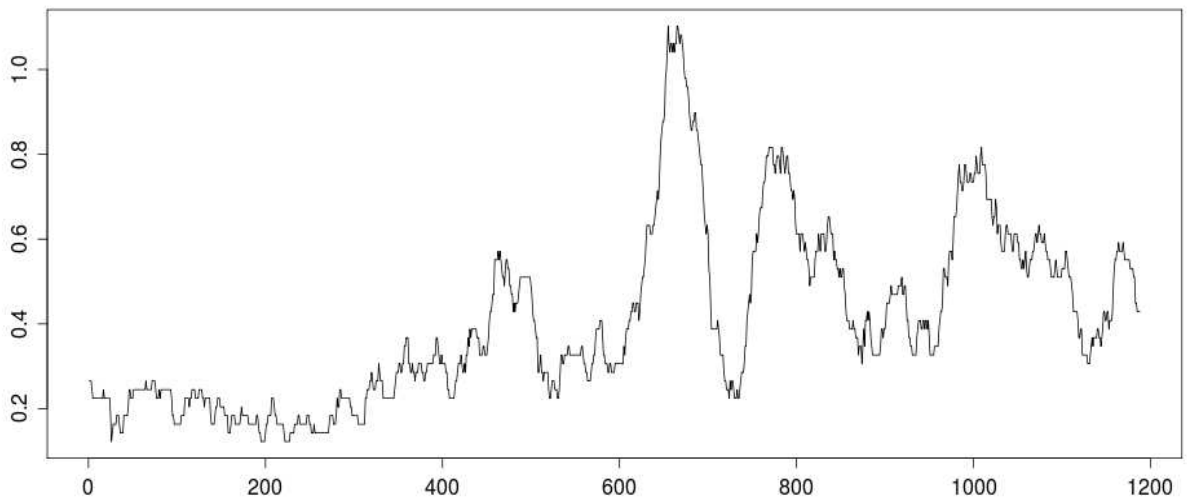


The study of temporal phenomena in history

Three levels of time in history (Braudel, *The Mediterranean*)

1. The geographical time (the time of the environment, with its slow, almost imperceptible change)
2. The social time (long-term social, economic, and cultural history)
3. The event time (the history of individuals with names : events, politics and people)

General tools : smoothing and highlighting trends (Labrousse)

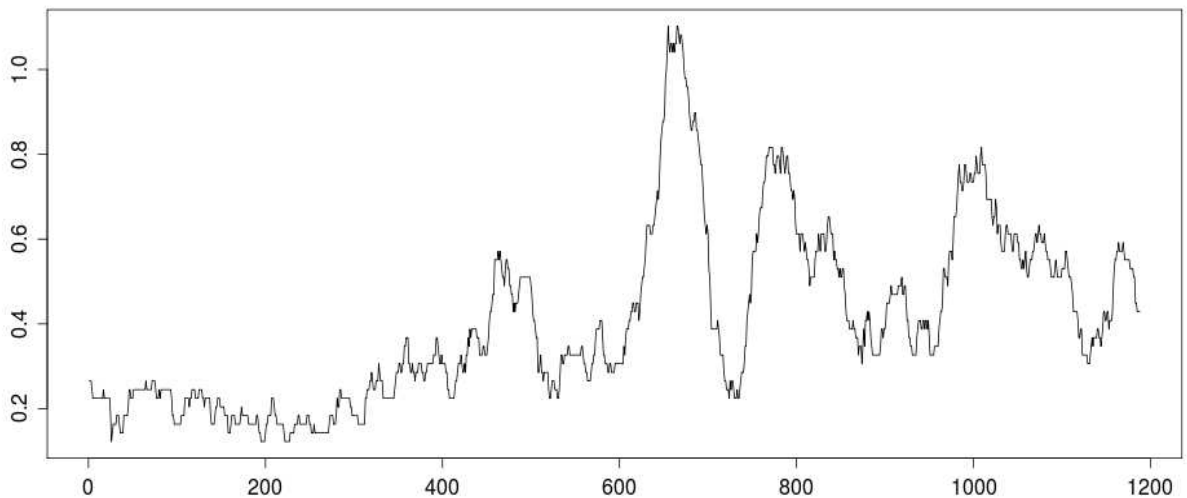


The study of temporal phenomena in history

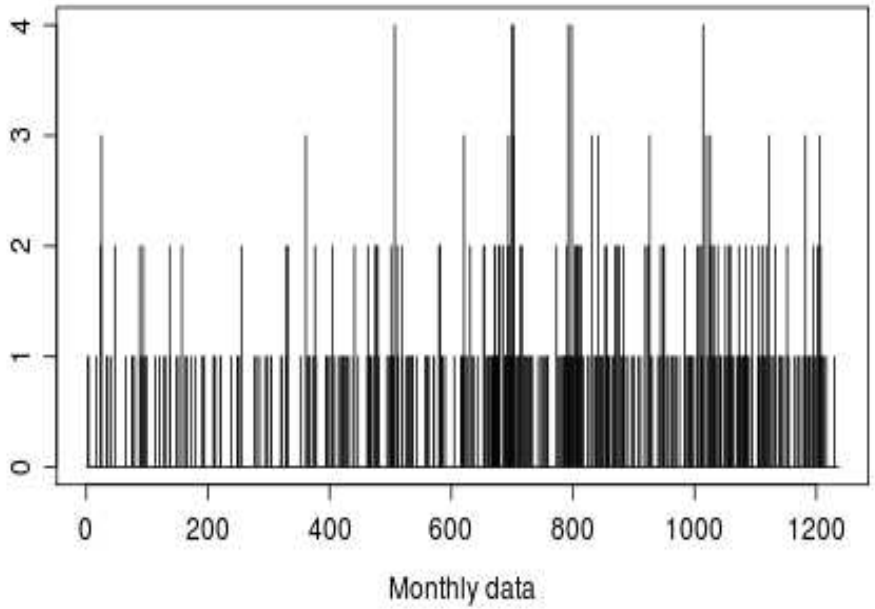
Three levels of time in history (Braudel, *The Mediterranean*)

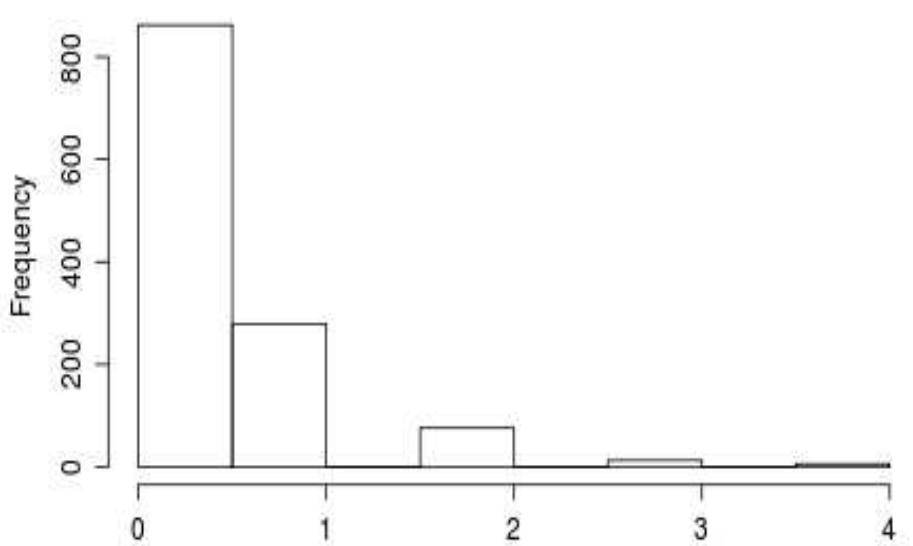
1. The geographical time (the time of the environment, with its slow, almost imperceptible change)
2. The social time (long-term social, economic, and cultural history)
3. The event time (the history of individuals with names : events, politics and people)

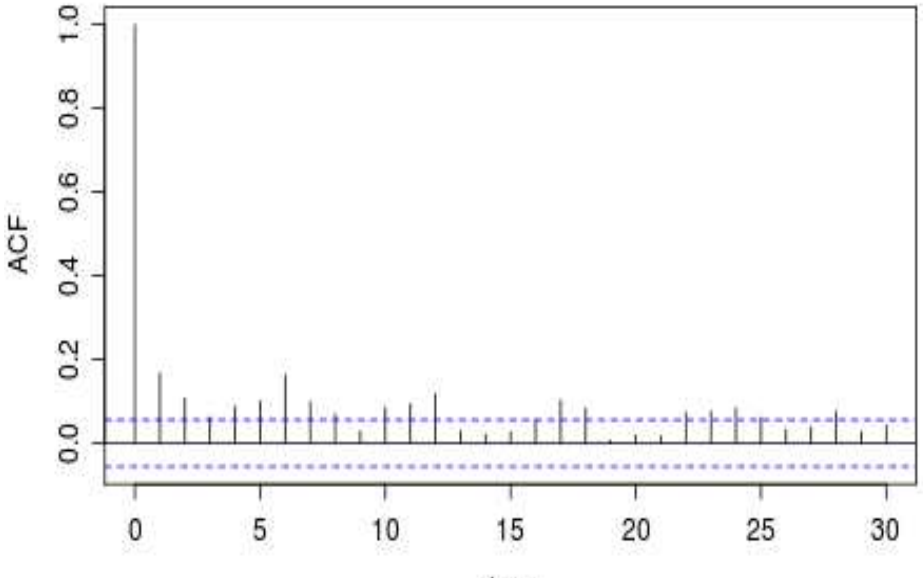
General tools : smoothing and highlighting trends (Labrousse)

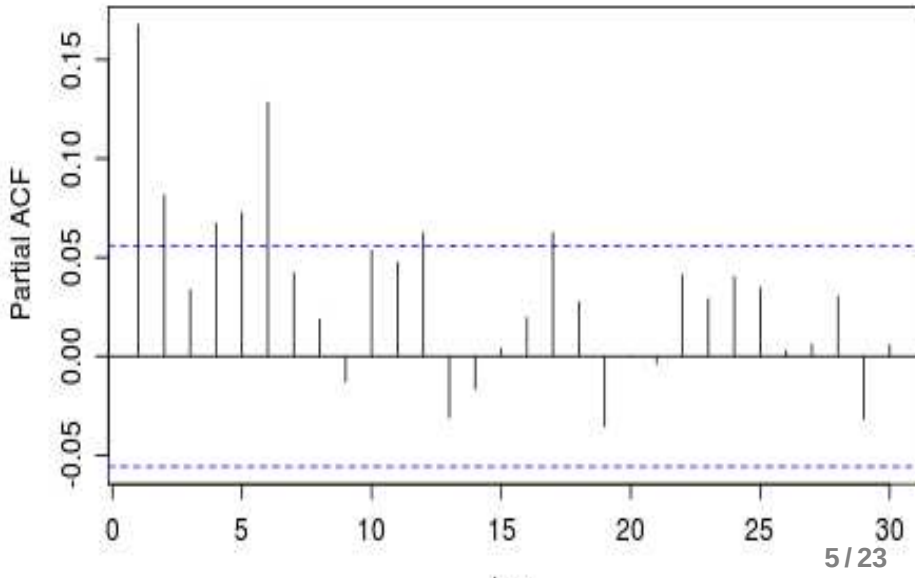


Monthly-issued logistics texts









Is there a link between issuing legislation and war periods ?

	Minimum	Mean	Maximum
War (45.7%)	0	0.54	4
Peace (54.3%)	0	0.28	4
Whole data-set	0	0.4	4

Welch test : t=6.27, df=1017.32, p-value=5.3e-10

Can we go further ?

- Is the State preparing for war ?
- Is the State subject to inertia after a war period ?

Is there a link between issuing legislation and war periods ?

	Minimum	Mean	Maximum
War (45.7%)	0	0.54	4
Peace (54.3%)	0	0.28	4
Whole data-set	0	0.4	4

Welch test : t=6.27, df=1017.32, p-value=5.3e-10

Can we go further ?

- Is the State preparing for war ?
- Is the State subject to inertia after a war period ?

1 Introduction

2 Time-series segmentation
- Zero-inflated Poisson - hidden Markov models (ZIP-HMM)
- INAR(p)-HMM model

3 Conclusion

Time-series segmentation

- Change-point detection (Lavielle, 2004)
- Regime-switching models
  - Hidden Markov chains (Baum et Petrie, 1966)
  - Autoregressive regime-switching models (Hamilton, 1990)
  - Hybrid models MLP-HMM (Rynkiewicz, 1999)
- Transitions or breaks ?
- Which models are adapted for integer-valued time series ?
- Which model take into account the over-dispersion in zero ?
- Which software ?

Time-series segmentation

- Change-point detection (Lavielle, 2004)
- Regime-switching models
  - Hidden Markov chains (Baum et Petrie, 1966)
  - Autoregressive regime-switching models (Hamilton, 1990)
  - Hybrid models MLP-HMM (Rynkiewicz, 1999)
- Transitions or breaks ?
- Which models are adapted for integer-valued time series ?
- Which model take into account the over-dispersion in zero ?
- Which software ?

First trip in Savoy : change-point detection with Matlab

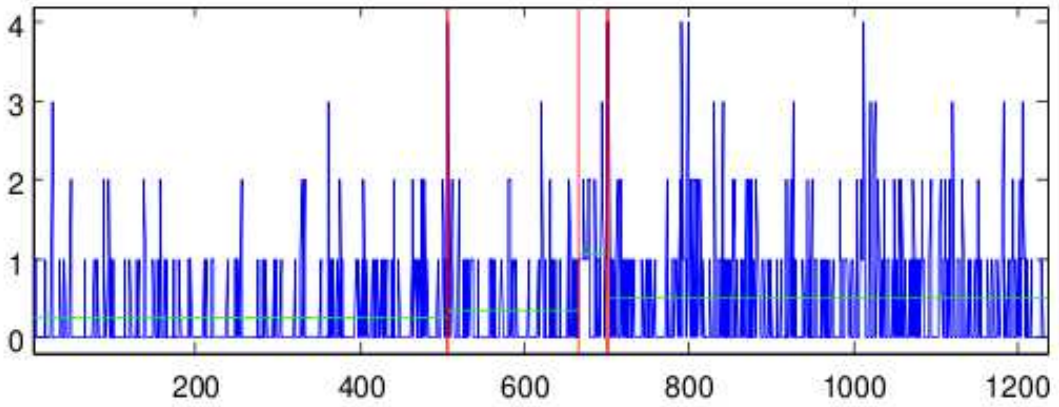


FIGURE: Change-points in mean

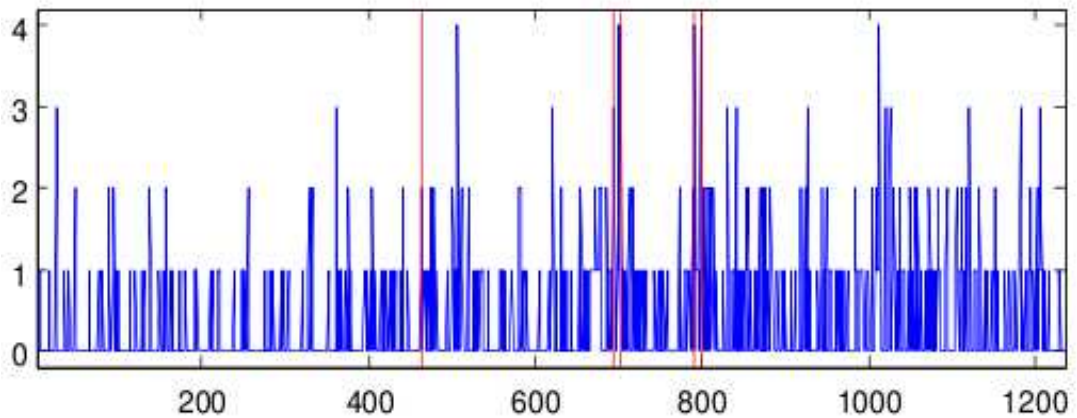


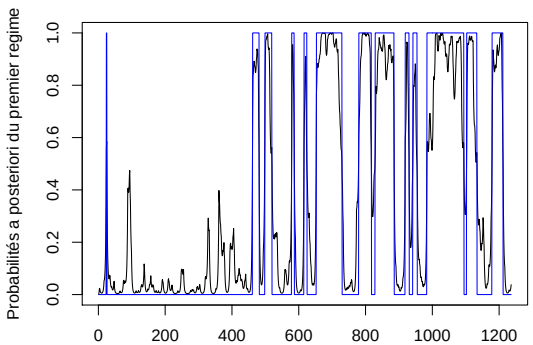
FIGURE: Change-points in variance

Second trip in Savoy : regime-switching models with R

- The first segmentation was performed with existing packages (HMM, HiddenMarkov)
- Only “classical“ distributions are implemented (Poisson, binomial, Gaussian,...)
- Poisson-HMM estimation

$$\pi = \begin{pmatrix} 0.97 & 0.03 \\ 0.05 & 0.95 \end{pmatrix}$$

$$\lambda = \begin{pmatrix} 0.20 \\ 0.73 \end{pmatrix}$$



- Overdispersion in zero ?
- Autoregressive structure ?

1 Introduction

2 Time-series segmentation
- Zero-inflated Poisson - hidden Markov models (ZIP-HMM)
- INAR(p)-HMM model

3 Conclusion

Hiddden-Markov models with zero-inflated Poisson distribution

- Zero-inflated Poisson distributions (Lambert, 1992)

$$(X_t) \sim \begin{cases} 0 & \text{with probability } \omega + (1-\omega)e^{-\lambda} \\ k > 0 & \text{with probability } (1-\omega)\frac{e^{-\lambda}\lambda^k}{k!} \end{cases}$$

- ZIP-HMM model

$$(X_t | S_t = e_i) \sim \begin{cases} 0 & \text{with probability } \omega_i + (1-\omega_i)e^{-\lambda_i} \\ k > 0 & \text{with probability } (1-\omega_i)\frac{e^{-\lambda_i}\lambda_i^k}{k!} \end{cases}$$

Hiddden-Markov models with zero-inflated Poisson distribution

- Zero-inflated Poisson distributions (Lambert, 1992)

$$(X_t) \sim \begin{cases} 0 & \text{with probability } \omega + (1-\omega)e^{-\lambda} \\ k > 0 & \text{with probability } (1-\omega)\frac{e^{-\lambda}\lambda^k}{k!} \end{cases}$$

- ZIP-HMM model

$$(X_t | S_t = e_i) \sim \begin{cases} 0 & \text{with probability } \omega_i + (1-\omega_i)e^{-\lambda_i} \\ k > 0 & \text{with probability } (1-\omega_i)\frac{e^{-\lambda_i}\lambda_i^k}{k!} \end{cases}$$

Parameter estimation

- The likelihood is maximized through the EM algorithm
- The complete likelihood may be written by introducing $Z_t|S_t = e_i \sim \mathcal{B}er(\omega_i)$

$$L(Z, X, S; \theta) = \prod_{t=1}^{T}\prod_{i=1}^{q} f(X_t, Z_t|S_t = e_i; \theta)^{\mathbf{1}_{e_i}(S_t)} \prod_{t=1}^{T-1}\prod_{i,j=1}^{q} \pi_{ij}^{\mathbf{1}_{e_i,e_j}(S_t,S_{t+1})} \times C$$

où $f(X_t, Z_t|S_t = e_i; \theta) = \omega_i^{\mathbf{1}_{(1,e_i)}(Z_t,S_t)}(1-\omega_i)^{\mathbf{1}_{(0,e_i)}(Z_t,S_t)} \left(\frac{e^{-\lambda_i}\lambda_i^{X_t}}{X_t!}\right)^{\mathbf{1}_{(0,e_i)}(Z_t,S_t)}$

E-step

$$
\begin{aligned}
Q(\theta|\theta^\star) = & \quad \mathbb{E}_{\theta^\star}\left[\ln(L(Z,X,S;\theta))|X_1^T\right] \\
= & \quad \sum_{t=1}^{T}\sum_{i=1}^{q}\{\mathbb{P}_{\theta^\star}(S_t=e_i,Z_t=1|X_1^T)\ln(\omega_i) \\
+ & \quad \mathbb{P}_{\theta^\star}(S_t=e_i,Z_t=0|X_1^T)(\ln(1-\omega_i)-\lambda_i+X_t\ln(\lambda_i)-\ln(X_t!))\} \\
+ & \quad \sum_{t=1}^{T}\sum_{i,j=1}^{q}\mathbb{P}_{\theta^\star}(S_{t-1}=e_i,S_t=e_j|X_1^T)\ln(\pi_{ij}) \qquad (1)
\end{aligned}
$$

$$
\mathbb{P}_{\theta^\star}(S_t=e_i,Z_t=1|X_1^T)=\mathbb{P}_{\theta^\star}(Z_t=1|X_1^T,S_t=e_i)\mathbb{P}_{\theta^\star}(S_t=e_i|X_1^T) \qquad (2)
$$

The last probability is computed with the Baum-Welch algorithm, while the first :

$$
\mathbb{P}_{\theta^\star}(Z_t=1|X_1^T,S_t=e_i)=\begin{cases}0 \text{ if } X_t>0\\ \frac{\mathbb{P}_{\theta^\star}(X_t=0|S_t=e_i,Z_t=1)\mathbb{P}_{\theta^\star}(Z_t=1|S_t=e_i)}{\mathbb{P}_{\theta^\star}(X_t=0|S_t=e_i)} \text{ if } X_t=0\end{cases}
$$

M-step

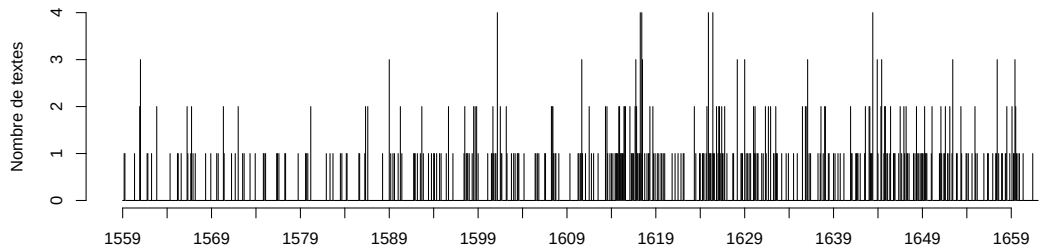
- Updates are computed analytically :

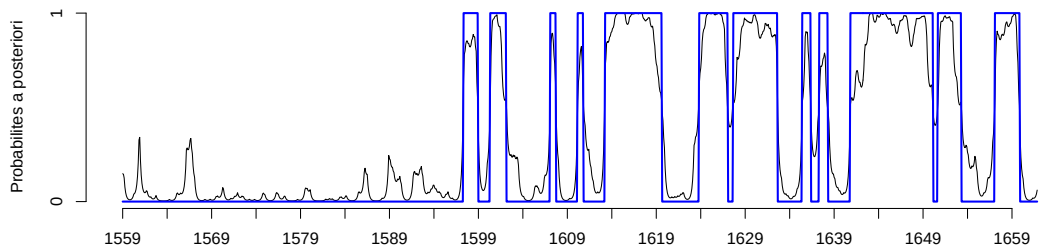
$$\pi_{ij} = \frac{\sum_{t=2}^{T} \mathbb{P}_{\theta^\star}(S_{t-1} = e_i, S_t = e_j | X_1^T)}{\sum_{t=1}^{T} \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)}$$

$$\omega_i = \frac{\alpha_i^\star \sum_{X_t=0} \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)}{\sum_{t:t=1}^{T} \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)}$$

$$\lambda_i = \frac{\sum_{t:X_t>0} \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T) * X_t}{\sum_{t:X_t=0} \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)(1-\alpha_i^\star) + \sum_{t:X_t>0} \mathbb{P}_{\theta^\star}(S_t = e_i | X_1^T)}$$

ZIP-HMM : the time of events





$$\pi = \begin{pmatrix} 0.98 & 0.02 \\ 0.04 & 0.96 \end{pmatrix} \quad \lambda = \begin{pmatrix} 0.30 \\ 0.77 \end{pmatrix} \quad \omega = \begin{pmatrix} 0.26 \\ 0.06 \end{pmatrix}$$

1 Introduction

2 Time-series segmentation
- Zero-inflated Poisson - hidden Markov models (ZIP-HMM)
- INAR(p)-HMM model

3 Conclusion

The INAR(p) model

- Take into account the dependence on the past
- The INAR(p) model (Al-Osh, 1987, 1990) :

$$X_t = \alpha_1 \circ X_{t-1} + \alpha_2 \circ X_{t-2} + \cdots \alpha_p \circ X_{t-p} + \epsilon_t \tag{3}$$

where $\epsilon_t \sim \mathcal{P}(\lambda)$ et $\alpha_i \circ X_{t-i} = \sum_{j=1}^{X_{t-i}} \xi_{ji}$ with $\xi_{ji} \sim \mathcal{B}er(\alpha_i)$ independent of ϵ_t and X_{t-i}.

- The conditional density

$$f(x_t | x_{t-1}, \cdots, x_{t-p}; \theta) =$$

$$\sum_{i_1=0}^{x_t \wedge x_{t-1}} C_{x_{t-1}}^{i_1} \alpha_1^{i_1} (1-\alpha_1)^{x_{t-1}-i_1} \sum_{i_2=0}^{x_t - i_1 \wedge x_{t-2}} C_{x_{t-2}}^{i_2} \alpha_2^{i_2} (1-\alpha_2)^{x_{t-2}-i_2} \cdots$$

$$\sum_{i_p=0}^{(x_t - i_1 - \cdots - i_{p-1}) \wedge x_{t-p}} C_{x_{t-p}}^{i_p} \alpha_2^{i_p} (1-\alpha_2)^{x_{t-p}-i_p} \frac{e^{-\lambda} \lambda^{x_t - i_1 - \cdots - i_p}}{(x_t - i_1 - \cdots - i_p)!}$$

The INAR(p)-HMM model

- The INAR parameters depend on the states of a unobserved Markov chain :

$$(X_t|S_t = e_i) = \alpha_{1,i} \circ X_{t-1} + \alpha_{2,i} \circ X_{t-2} + \cdots \alpha_{p,i} \circ X_{t-p} + \epsilon_{i,t} \tag{4}$$

  where $\epsilon_{i,t} \sim \mathcal{P}(\lambda_i)$

- The log-likelihood is maximized via the EM algorithm
- The complete conditional likelihood :

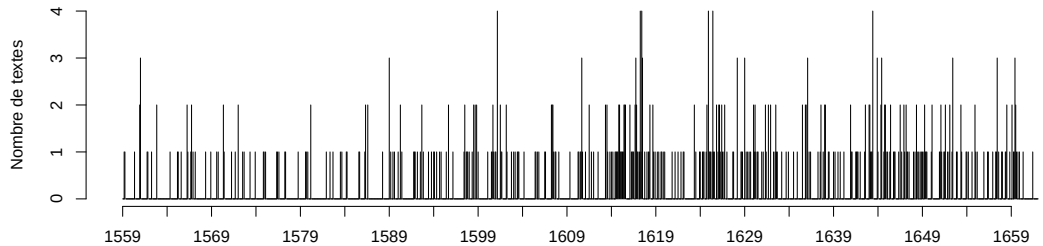
$$L_n(X_p^n, S_1^n|X_1^p; \theta) = \prod_{k=p}^{n} \pi(S_{k-1}; S_k) \prod_{k=p}^{n} f(X_k|X_{k-1}, \cdots, X_{k-p}, S_k; \theta) Cte \tag{5}$$

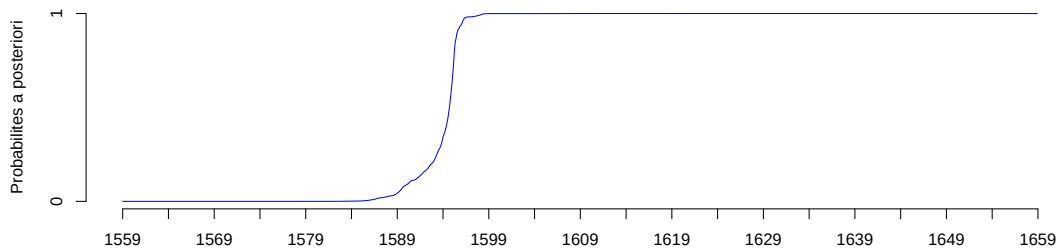
- E-step :

$$\mathbb{E}_{\theta'}\{log L_n(X_1^n, S_1^n; \theta)|X_1^n\} = \sum_{k=p}^{n} \sum_{s,s'} log(\pi(S_{k-1} = s; S_k = s'))\mathbb{P}_{\theta'}(S_{k-1} = s, S_k = s'|X_1^n)$$
$$+ \sum_{k=p}^{n} \sum_{s} log(f(X_k|X_{k-1}, \cdots, X_{k-p}, S_k = s; \theta))\mathbb{P}_{\theta'}(S_k = s|X_1^n)$$

- M-step is computed numerically

INAR(p)-HMM : the social time





$$\pi = \begin{pmatrix} 0.998 & 0.002 \\ 0 & 1 \end{pmatrix} \quad \lambda = \begin{pmatrix} 0.205 \\ 0.306 \end{pmatrix}$$

$$\alpha = \begin{pmatrix} 0.0406 & 0.0041 & 0.0001 & 0.0001 & 0.0001 & 0.0001 \\ 0.114 & 0.0617 & 0.0011 & 0.046 & 0.0521 & 0.117 \end{pmatrix}$$

Conclusion

- Time-series segmentation on historical data is achieved on various time-scales
- Humanities & social sciences data requires adapted methods (integer-valued, over-dispersion in zero, irregular time series, fuzzy or imprecise values)
- DiscreteTS may be used complementary to HMM or HidenMarkov

References

[1] Braudel F. (1949) La Méditerranée et le monde méditerranéen à l'époque de Philippe II, Paris, Armand Colin
[2] http ://cran.r-project.org/web/packages/HiddenMarkov/index.html
[3] http ://cran.r-project.org/web/packages/HMM/index.html
[3] Olteanu M., Ridgway J. (2012). Hidden Markov models for time series of counts with excess zeros. *Proceedings of ESANN 2012*, 133-138
[4] Ridgway J. (2011). Hidden Markov models for time series of count data. *Rapport de stage*