A comparisons of some criteria for states selection of the latent Markov model for longitudinal data

Silvia Bacci*[1], Francesco Bartolucci*, Silvia Pandolfi*, Fulvia Pennoni**

*Dipartimento di Economia, Finanza e Statistica - Università di Perugia
**Dipartimento di Statistica - Università di Milano-Bicocca

Università di Catania, Catania, 6-7 September 2012

[1] silvia.bacci@stat.unipg.it

Outline

Introduction

- Background:

  Latent Markov (LM) models (Wiggins, 1973; Bartolucci et al., 2012) are successfully applied in the analysis of longitudinal data: they allow to take into account several aspects, such as serial dependence between observations, measurement errors, unobservable heterogeneity

  LM models assume that one or more occasion-specific response variables depends only on a discrete latent variable characterized by a given number of latent states which in turn depends on the latent variables corresponding to the previous occasions according to a first-order Markov chain

  LM models are characterized by several parameters: the initial probabilities to belong to a given latent state, the transition probabilities from a latent state to another one, the conditional response probabilities given the discrete latent variable

- Problem: a crucial point with LM models is represented by the selection of the number of latent states

- Aim: we compare the behavior of several model selection criteria to choose the number of latent states

  Special attention is devoted to classification-based criteria that take explicitly into account the partition of observations in different latent states, through a specific measurement of the quality of classification, denoted as entropy

Multivariate basic LM model: notation

- $\boldsymbol{Y}^{(t)} = (Y_1^{(t)}, \ldots, Y_r^{(t)})$: vector of discrete categorical response variables Y_j ($j = 1, \ldots, r$) observed at time t ($t = 1, \ldots, T$), having c_j categories
- $\boldsymbol{Y} = (\boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(T)})$: vector of observed responses made of the union of vectors $\boldsymbol{Y}^{(t)}$; usually, it is referred to repeated measurements of the same variables Y_j ($j = 1, \ldots, r$) on the same individuals at different time points
- $U^{(t)}$: latent state at time t with state space $\{1, \ldots, k\}$
- $\boldsymbol{U} = (U^{(1)}, \ldots, U^{(T)})$: vector describing the latent process

Multivariate basic LM model: main assumptions

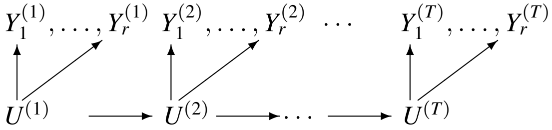
- vectors $\boldsymbol{Y}^{(t)}$ $(t = 1, \ldots, T)$ are conditionally independent given the latent process $\boldsymbol{U}$ and the response variables in each $\boldsymbol{Y}^{(t)}$ are conditionally independent given $U^{(t)}$ (local independence), i.e.,

  each occasion-specific observed variable $Y_j^{(t)}$ is independent of $Y_j^{(t-1)}, \ldots, Y_j^{(1)}$ and of each $Y_h^{(t)}$, for all $h \neq j = 1, \ldots, r$, given $U^{(t)}$

- latent process $\boldsymbol{U}$ follows a first-order Markov chain with k latent states, i.e.,

  each latent variable $U^{(t)}$ is independent of $U^{(t-2)}, \ldots, U^{(1)}$, given $U^{(t-1)}$

LM:



Multivariate basic LM model: parameters

- $k\sum_{j=1}^r (c_j-1)$ conditional response probabilities

  $\phi_{jy|u}^{(t)} = p(Y_j^{(t)} = y|U^{(t)} = u) \quad j = 1,\dots,r;\ t = 1,\dots,T;\ u = 1,\dots,k;\ y = 0,\dots,c_j-1$

  $\phi_{\boldsymbol{y}|u}^{(t)} = \prod_{j=1}^r \phi_{jy|u}^{(t)} = p(Y_1^{(t)} = y_1,\dots,Y_r^{(t)} = y_r|U^{(t)} = u)$

- $(k-1)$ initial probabilities

  $\pi_u = p(U^{(1)} = u) \quad u = 1,\dots,k$

- $(T-1)k(k-1)$ transition probabilities

  $\pi_{u|v}^{(t|t-1)} = p(U^{(t)} = u|U^{(t-1)} = v) \quad t = 2,\dots,T;\ u,v = 1,\dots,k$

- $\#\text{par} = k\sum_{j=1}^r (c_j-1) + (k-1) + (T-1)k(k-1)$

Multivariate basic LM model: probability distributions

- $p(\boldsymbol{U} = \boldsymbol{u}) = \pi_u \prod_{t=2}^T \pi_{u|v}^{(t|t-1)} = \pi_u \cdot \pi_{u_2|u}^{(2|1)} \dots \pi_{u_T|u_{T-1}}^{(T|T-1)}$
- $p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{U} = \boldsymbol{u}) = \prod_{t=1}^T \phi_{\boldsymbol{y}|u}^{(t)} = \phi_{\boldsymbol{y}|u}^{(1)} \cdot \phi_{\boldsymbol{y}|u}^{(2)} \dots \phi_{\boldsymbol{y}|u}^{(T)}$
- manifest distribution of $\boldsymbol{Y}$

$$p(\boldsymbol{Y} = \boldsymbol{y}) = \sum_{\boldsymbol{u}} p(\boldsymbol{Y} = \boldsymbol{y}, \boldsymbol{U} = \boldsymbol{u}) = \sum_{\boldsymbol{u}} p(\boldsymbol{U} = \boldsymbol{u}) \cdot p(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{U} = \boldsymbol{u})$$
$$= \sum_{u} \pi_u \phi_{\boldsymbol{y}|u}^{(1)} \cdot \sum_{u_2} \pi_{u_2|u}^{(2|1)} \phi_{\boldsymbol{y}|u}^{(2)} \dots \sum_{u_T} \pi_{u_T|u_{T-1}}^{(T|T-1)} \phi_{\boldsymbol{y}|u}^{(T)}$$
$$= \sum_{u} \sum_{u_2} \dots \sum_{u_T} \pi_u \prod_{t=2}^{T} \pi_{u|v}^{(t|t-1)} \prod_{t=1}^{T} \phi_{\boldsymbol{y}|u}^{(t)}$$

Note that computing $p(\boldsymbol{Y} = \boldsymbol{y})$ involves all the possible k^T configurations of vector $\boldsymbol{u}$

Multivariate basic LM model: maximum likelihood (ML) estimation

- Log-likelihood of the model

$$\ell(\boldsymbol{\theta}) = \sum_{\boldsymbol{y}} n_{(\boldsymbol{y})} \log[p(\boldsymbol{Y} = \boldsymbol{y})]$$

  - $\boldsymbol{\theta}$: vector of all model parameters (π_u, $\pi_{u|v}^{(t|t-1)}$, $\phi_{jy|u}^{(t)}$)
  - $n_{(\boldsymbol{y})}$: frequency of the response configuration $\boldsymbol{y}$ in the sample

- $\ell(\boldsymbol{\theta})$ may be maximized with respect to $\boldsymbol{\theta}$ by an *Expectation-Maximization (EM) algorithm* (Dempster *et al.*, 1977)

EM algorithm

Complete data log-likelihood of the model

$$\ell^*(\boldsymbol{\theta}) = \sum_{j=1}^{r}\sum_{t=1}^{T}\sum_{u=1}^{k}\sum_{y=0}^{c-1} a_{juy}^{(t)} \log \phi_{jy|u}^{(t)} + \\ + \sum_{u=1}^{k} b_u^{(1)} \log \pi_u + \sum_{t=2}^{T}\sum_{v=1}^{k}\sum_{u=1}^{k} b_{vu}^{(t)} \log \pi_{u|v}^{(t|t-1)}$$

- $a_{juy}^{(t)}$: frequency of subjects responding by y for the j-th response variable and belonging to latent state u, at time t
- $b_u^{(1)}$: frequency of subjects in latent state u at time 1
- $b_{vu}^{(t)}$: frequency of subjects which move from latent state v to u at time t

EM algorithm

- The algorithm *alternates two steps* until convergence in $\ell(\boldsymbol{\theta})$:

  E: compute the expected values of frequencies $a_{juy}^{(t)}$, $b_u^{(1)}$, and $b_{vu}^{(t)}$, given the observed data and the current value of $\boldsymbol{\theta}$, so as to obtain the expected value of $\ell^*(\boldsymbol{\theta})$

  M: update $\boldsymbol{\theta}$ by maximizing the expected value of $\ell^*(\boldsymbol{\theta})$ obtained above; explicit solutions for $\boldsymbol{\theta}$ estimations are available

- The E-step is performed by means of certain recursions which may be easily implemented through matrix notation (Bartolucci, 2006)

Forward and backward recursions

To efficiently compute the probability $p(\boldsymbol{Y} = \boldsymbol{y})$ and the posterior probabilities $f_{u|\boldsymbol{y}}^{(t)}$ and $f_{u|v,\boldsymbol{y}}^{(t|t-1)}$ we can use forward and backward recursions for obtaining the following intermediate quantities

- Forward recursions

$$q_{u,\boldsymbol{y}}^{(t)} = p(U^{(t)} = u, \boldsymbol{Y}^{(1)}, \ldots, \boldsymbol{Y}^{(t)}) = \sum_{v=1}^{k} q_{v,\boldsymbol{y}}^{(t-1)} \pi_{u|v}^{(t|t-1)} \phi_{\boldsymbol{y}|u}^{(t)} \quad u = 1, \ldots, k$$

starting with $q_{u,\boldsymbol{y}}^{(1)} = \pi_u \phi_{\boldsymbol{y}|u}^{(1)}$

- Backward recursions

$$\bar{q}_{v,\boldsymbol{y}}^{(t)} = p(\boldsymbol{Y}^{(t+1)}, \ldots, \boldsymbol{Y}^{(T)} | U^{(t)} = v) = \sum_{u=1}^{k} \bar{q}_{u,\boldsymbol{y}}^{(t+1)} \pi_{u|v}^{(t+1|t)} \phi_{\boldsymbol{y}|u}^{(t+1)} \quad v = 1, \ldots, k$$

starting with $\bar{q}_{v,\boldsymbol{y}}^{(T)} = 1$

Model selection criteria

- A crucial point with LM models concerns the selection of k, the number of latent states
- We may rely on the literature about finite mixture models and hidden Markov models
- The most well-known criteria are
  - Akaike's Information Criterion (AIC - Akaike, 1973)

$$\mathrm{AIC} = -2\ell(\boldsymbol{\theta}) + 2 \cdot \#\mathrm{par}$$

    or its variants:
  - Consistent AIC (CAIC)

$$\mathrm{CAIC} = -2\ell(\boldsymbol{\theta}) + \#\mathrm{par} \cdot (\log(n) + 1)$$

  - AIC3

$$\mathrm{AIC3} = -2\ell(\boldsymbol{\theta}) + 3 \cdot \#\mathrm{par}$$

  - Bayesian Information Criterion (BIC - Schwarz, 1978)

$$\mathrm{BIC} = -2\ell(\boldsymbol{\theta}) + \#\mathrm{par} \cdot \log(n)$$

Classification-based criteria

Some criteria are developed in the context of the classification likelihood approach, based on the relation

$$\ell^*(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}) - \text{EN}$$

where EN is the entropy and it denotes a penalization term which measures the quality of the partition and it is defined as (Hernando et al., 2005)

$$\begin{aligned}\text{EN} &= -\sum_{u_1}\ldots\sum_{u_T} f_{u_1,\ldots u_T|\boldsymbol{y}} \log(f_{u_1,\ldots u_T|\boldsymbol{y}}) = \\ &= -\sum_{u_1}\ldots\sum_{u_T} f^{(1)}_{u_1|\boldsymbol{y}} \cdot f^{(2|1)}_{u_2|u_1,\boldsymbol{y}} \cdot \ldots \cdot f^{(t|t-1)}_{u_t|u_{t-1},\boldsymbol{y}} \cdot \ldots \cdot f^{(T|T-1)}_{u_T|u_{T-1},\boldsymbol{y}} \cdot \\ &\cdot [\log(f^{(1)}_{u_1|\boldsymbol{y}}) + \log(f^{(2|1)}_{u_2|u_1,\boldsymbol{y}}) + \ldots + \log(f^{(t|t-1)}_{u_t|u_{t-1},\boldsymbol{y}}) + \ldots + \log(f^{(T|T-1)}_{u_T|u_{T-1},\boldsymbol{y}})]\end{aligned}$$

with

$$f_{u|\boldsymbol{y}}^{(t)} = \frac{q_{u,\boldsymbol{y}}^{(t)} \cdot \bar{q}_{u,\boldsymbol{y}}^{(t)}}{p(\boldsymbol{Y} = \boldsymbol{y})}$$

$$f_{u|v,\boldsymbol{y}}^{(t|t-1)} = \frac{f_{v,u|\boldsymbol{y}}^{(t-1,t)}}{f_{v|\boldsymbol{y}}^{(t-1)}} = \frac{q_{v,\boldsymbol{y}}^{(t-1)} \pi_{u|v}^{(t|t-1)} \phi_{\boldsymbol{y}^{(t)}|u} \bar{q}_{u,\boldsymbol{y}}^{(t)}}{p(\boldsymbol{Y} = \boldsymbol{y})} \cdot \frac{p(\boldsymbol{Y} = \boldsymbol{y})}{q_{v,\boldsymbol{y}}^{(t-1)} \bar{q}_{v,\boldsymbol{y}}^{(t-1)}} =$$

$$= \pi_{u|v}^{(t|t-1)} \phi_{\boldsymbol{y}^{(t)}|u} \cdot \frac{\bar{q}_{u,\boldsymbol{y}}^{(t)}}{\bar{q}_{v,\boldsymbol{y}}^{(t-1)}}$$

We may also formulate an approximation for EN, under the assumption that $u^{(t)}$ are independent given $\boldsymbol{Y}$:

- $\mathrm{EN}_1 = -\sum_{u_1}\dots\sum_{u_T} f_{u|\mathbf{y}}^{(t)} \log(f_{u|\mathbf{y}}^{(t)})$
- or a possible variant of EN_1 given by $\mathrm{EN}_2 = -\sum_{u_1}\dots\sum_{u_T} f_{u|\mathbf{y}}^{(t)} \log(f_{u|\mathbf{y}}^{(t)})/T$
- Example: T=3

$$
\begin{aligned}
\mathrm{EN} &= -\sum_u\sum_v\sum_z f_{u,v,z|\mathbf{y}} \log(f_{u,v,z|\mathbf{y}}) = \\
&= f_{z|v,\mathbf{y}}^{(3|2)} \cdot f_{v|u,\mathbf{y}}^{(2|1)} \cdot f_{u|\mathbf{y}}^{(1)} \cdot \\
&\cdot [\log(f_{z|v,\mathbf{y}}^{(3|2)}) + \log(f_{v|u,\mathbf{y}}^{(2|1)}) + \log(f_{u|\mathbf{y}}^{(1)})] \\
\mathrm{EN}_1 &= -[f_{u|\mathbf{y}}^{(1)} \cdot \log(f_{u|\mathbf{y}}^{(1)}) + f_{v|\mathbf{y}}^{(2)} \cdot \log(f_{v|\mathbf{y}}^{(2)}) + f_{z|\mathbf{y}}^{(3)} \cdot \log(f_{z|\mathbf{y}}^{(3)})] \\
\mathrm{EN}_2 &= \frac{1}{3}\mathrm{EN}_1
\end{aligned}
$$

Some classification-based criteria are (McLachlan and Peel, Chap. 6)

- Classification Likelihood information Criterion (CLC)

$$\mathrm{CLC} = -2\ell(\boldsymbol{\theta}) + 2 \cdot \mathrm{EN}$$

- Approximated Integrated Classification Likelihood criterion (ICL-BIC)

$$\mathrm{ICL} - \mathrm{BIC} = \mathrm{BIC} + 2 \cdot \mathrm{EN}$$

- Normalized Entropy Criterion (NEC)

$$\mathrm{NEC} = \frac{\mathrm{EN}}{\ell(\boldsymbol{\theta}) - \ell_1(\boldsymbol{\theta})} \quad k \geq 2$$

  where $\ell_1(\boldsymbol{\theta})$ is the maximum log-likelihood in case of $k = 1$, and $\mathrm{NEC} = 1$ if $k = 1$

- Approximated NECs:

$$\mathrm{NEC}_1 = \frac{\mathrm{EN}_1}{\ell(\boldsymbol{\theta}) - \ell_1(\boldsymbol{\theta})} \quad k \geq 2$$

$$\mathrm{NEC}_2 = \frac{\mathrm{EN}_2}{\ell(\boldsymbol{\theta}) - \ell_1(\boldsymbol{\theta})} \quad k \geq 2$$

Monte Carlo simulation study

- We compare
  - AIC, CAIC, AIC3, BIC
  - CLC, ICL-BIC, NEC, NEC_1, NEC_2
- 100 samples with a given size n and coming from a multivariate LM model, characterized by r binary $(y = 0, 1)$ response variables observed in T time occasions, k latent states, and given values of initial probabilities π_u, transition probabilities $\pi_{u|v}^{(t|t-1)}$, conditional response probabilities $\phi_{jy|u}^{(t)}$
- $n = 250, 500, 1000$
- $r = 1, 3, 5$
- $T = 5, 10$
- $k = 2, 3$
- all analyses are implemented in R software

Main results

Scenery 1

- $n = 250$, $T = 5$, $k = 2$
- $\phi^{(t)}_{j0|u=1} = 0.8 = \phi^{(t)}_{j1|u=2}$, $\quad \phi^{(t)}_{j0|u=2} = 0.2 = \phi^{(t)}_{j1|u=1}$
- $\pi_1 = 0.5 = \pi_2$
- $\pi^{(t|t-1)}_{1|1} = 0.7 = \pi^{(t|t-1)}_{2|2}$, $\quad \pi^{(t|t-1)}_{1|2} = 0.3 = \pi^{(t|t-1)}_{2|1}$ (time homogenous assumption)
- $r = 1, 3, 5$

Main results

Scenery 1: Relative frequencies of k chosen on the basis of several criteria

k	BIC	AIC	AIC3	CAIC	NEC	NEC_1	NEC_2	CLC	ICL-BIC
$r = 1$									
1	**0.52**	0.00	0.10	**0.63**	**1.00**	**1.00**	**0.99**	**1.00**	**1.00**
2	**0.48**	0.98	0.90	0.37	0.00	0.00	0.01	0.00	0.00
3	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	**0.88**	**0.92**	0.00	**0.88**	**0.95**
2	1.00	0.83	0.98	1.00	0.10	0.07	0.96	0.10	0.04
3	0.00	0.16	0.02	0.00	0.01	0.01	0.04	0.01	0.01
4	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	1.00	0.77	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Main results

Scenery 2

- $n = 250$, $T = 5$, $k = 2$
- $\phi^{(t)}_{j0|u=1} = 0.7 = \phi^{(t)}_{j1|u=2}, \quad \phi^{(t)}_{j0|u=2} = 0.3 = \phi^{(t)}_{j1|u=1}$
- $\pi_1 = 0.5 = \pi_2$
- $\pi^{(t|t-1)}_{1|1} = 0.9 = \pi^{(t|t-1)}_{2|2}, \quad \pi^{(t|t-1)}_{1|2} = 0.1 = \pi^{(t|t-1)}_{2|1}$ (time homogenous assumption)
- $r = 1, 3, 5$

Main results

Scenery 2: Relative frequencies of k chosen on the basis of several criteria

k	BIC	AIC	AIC3	CAIC	NEC	NEC_1	NEC_2	CLC	ICL-BIC
$r = 1$									
1	0.35	0.01	0.02	**0.53**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
2	0.65	0.98	0.97	**0.47**	0.00	0.00	0.00	0.00	0.00
3	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
$r = 3$									
1	0.00	0.00	0.00	0.00	**1.00**	**1.00**	0.09	**1.00**	**1.00**
2	1.00	0.92	0.995	1.00	0.00	0.00	0.855	0.00	0.00
3	0.00	0.07	0.005	0.00	0.00	0.00	0.015	0.00	0.00
4	0.00	0.01	0.00	0.00	0.00	0.00	0.015	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00	0.00	0.025	0.00	0.00
$r = 5$									
1	0.00	0.00	0.00	0.00	0.285	**0.77**	0.00	0.285	**0.55**
2	1.00	0.78	0.995	1.00	0.59	0.22	0.98	0.59	**0.445**
3	0.00	0.205	0.005	0.00	0.03	0.005	0.015	0.035	0.005
4	0.00	0.01	0.00	0.00	0.07	0.005	0.005	0.070	0.00
5	0.00	0.005	0.00	0.00	0.025	0.00	0.000	0.025	0.00

Main results

Scenery 3

- $n = 500$, $T = 5$, $k = 3$
- $\phi^{(t)}_{j0|u=1} = 0.9 = \phi^{(t)}_{j1|u=2}$, $\phi^{(t)}_{j0|u=2} = 0.1 = \phi^{(t)}_{j1|u=1}$, $\phi^{(t)}_{j0|u=3} = 0.4$, $\phi^{(t)}_{j1|u=3} = 0.6$
- $\pi_1 = \pi_2 = \pi_3 = 0.33$
- $\pi^{(t|t-1)}_{1|1} = \pi^{(t|t-1)}_{2|2} = \pi^{(t|t-1)}_{3|3} = 0.80$, $\pi^{(t|t-1)}_{2|1} = 0.15 = \pi^{(t|t-1)}_{2|3}$, $\pi^{(t|t-1)}_{3|1} = 0.05 = \pi^{(t|t-1)}_{1|3}$, $\pi^{(t|t-1)}_{1|2} = 0.10 = \pi^{(t|t-1)}_{3|2}$ (time homogenous assumption)
- $r = 1, 3, 5$

Main results

Scenery 3: Relative frequencies of k chosen on the basis of several criteria

k	BIC	AIC	AIC3	CAIC	NEC	NEC_1	NEC_2	CLC	ICL-BIC
$r=1$									
1	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**0.92**	**1.00**	**1.00**
2	**1.00**	**0.98**	**0.99**	**1.00**	0.00	0.00	0.07	0.00	0.00
3	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00	0.00
$r=3$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.03	0.00	0.00	0.10	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
3	0.97	0.81	1.00	0.90	0.00	0.00	0.00	0.00	0.00
4	0.00	0.19	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$r=5$									
1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
3	1.00	0.78	0.99	1.00	0.00	0.00	0.00	0.00	0.00
4	0.00	0.20	0.01	0.00	0.00	0.00	0.00	0.00	0.00
5	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Conclusions

- We compared several criteria for the selection of the number of latent states in the LM models
- We observed that:
  - AIC, BIC and their variants present a better general behavior with respect to the classification-based criteria
  - classification-based criteria tend to underestimate the true number of latent states, mainly for the univariate case
  - the behavior of classification-based criteria improves by increasing the number of observed response variables
  - by increasing the number k of latent states the performance of all considered criteria gets worse
- For further developments of our work, we would like to study in deep extended versions of entropy and classification-based criteria to improve the performance of the latent states selection process
- We will refer to the most recent developments in the context of hidden Markov models: see Durand and Guedon (2012) for a discussion about the tendency of entropy to overestimate the uncertainty and for a new proposal to decompose the global entropy in conditional entropies

Main references

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, Second International symposium of information theory, pages 267-281, Budapest. Akademiai Kiado.

Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. Journal of the Royal Statistical Society, series B, 68:155Ð178.

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2012), Latent Markov Models for longitudinal data: Applications in Social Science and Economics, Chapman & Hall

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society, Series B, 39:1Ð38.

Durand, J-.B., Guédon, Y. (2012). Localizing the latent structure canonical uncertainty: entropy profiles for hidden Markov models, Research Report 7896, Project-Teams Mistis and Virtual Plants.

Hernando, D., Crespi, V., and Cybenko, G. (2005). Efficient computation of the hidden Markov model entropy for a given observation sequence. IEEE Transactions on Information Theory, 51(7), 2681-2685.

McLachlan, G. and Peel, D. (2000). Finite Mixture Models. Chap. 6. Wiley.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, 6, 461-464.

Wiggins, L. (1973). Panel Analysis: Latent probability models for attitude and behaviours processes. Elsevier, Amsterdam.