Learning from data in Markov models

Marco Cattaneo
Department of Statistics, LMU Munich
cattaneo@stat.uni-muenchen.de

WPMSIIP 2011, Ljubljana, Slovenia
10 September 2011

Markov chains

- a Markov chain $X_1, X_2, \ldots$ with a finite set $\mathcal{S} = \{s_1, \ldots, s_k\}$ of possible states is described by its $k \times k$ transition matrices M_n (and by the distribution of X_1), where

$$M_{n,ij} = P(X_{n+1} = s_j \mid X_n = s_i)$$

Markov chains

- a Markov chain $X_1, X_2, \ldots$ with a finite set $\mathcal{S} = \{s_1, \ldots, s_k\}$ of possible states is described by its $k \times k$ transition matrices M_n (and by the distribution of X_1), where

$$M_{n,ij} = P(X_{n+1} = s_j \mid X_n = s_i)$$

- the Markov chain is homogeneous if $M_n = M$ does not depend on n

Markov chains

- a Markov chain $X_1, X_2, \ldots$ with a finite set $\mathcal{S} = \{s_1, \ldots, s_k\}$ of possible states is described by its $k \times k$ transition matrices M_n (and by the distribution of X_1), where

$$M_{n,ij} = P(X_{n+1} = s_j \,|\, X_n = s_i)$$

- the Markov chain is homogeneous if $M_n = M$ does not depend on n
- a (homogeneous) imprecise Markov chain has the transition matrix M replaced by a set $\mathcal{M}$ of transition matrices, and can be interpreted in (at least) two different ways:

Markov chains

- a Markov chain $X_1, X_2, \dots$ with a finite set $\mathcal{S} = \{s_1, \dots, s_k\}$ of possible states is described by its $k \times k$ transition matrices M_n (and by the distribution of X_1), where

$$M_{n,ij} = P(X_{n+1} = s_j \,|\, X_n = s_i)$$

- the Markov chain is homogeneous if $M_n = M$ does not depend on n
- a (homogeneous) imprecise Markov chain has the transition matrix M replaced by a set $\mathcal{M}$ of transition matrices, and can be interpreted in (at least) two different ways:
    - a homogeneous (precise) Markov chain, for which we only know that $M \in \mathcal{M}$

Markov chains

- a Markov chain $X_1, X_2, \dots$ with a finite set $\mathcal{S} = \{s_1, \dots, s_k\}$ of possible states is described by its $k \times k$ transition matrices M_n (and by the distribution of X_1), where

$$M_{n,ij} = P(X_{n+1} = s_j \,|\, X_n = s_i)$$

- the Markov chain is homogeneous if $M_n = M$ does not depend on n
- a (homogeneous) imprecise Markov chain has the transition matrix M replaced by a set $\mathcal{M}$ of transition matrices, and can be interpreted in (at least) two different ways:
  - a homogeneous (precise) Markov chain, for which we only know that $M \in \mathcal{M}$
  - an inhomogeneous (precise) Markov chain, for which we only know that $M_n \in \mathcal{M}$

learning from data

- assume that we observe some states $X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n}$ and we consider them as realizations from:

learning from data

- assume that we observe some states $X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n}$ and we consider them as realizations from:
  - a homogeneous (precise) Markov chain: then we learn something about the transition matrix M

learning from data

- assume that we observe some states $X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n}$ and we consider them as realizations from:
  - a homogeneous (precise) Markov chain: then we learn something about the transition matrix M
  - an inhomogeneous (precise) Markov chain (without additional assumptions): then we do not learn (almost) anything about the transition matrices M_n

learning from data

- assume that we observe some states $X_{i_1} = x_{i_1}, \dots, X_{i_n} = x_{i_n}$ and we consider them as realizations from:
    - a homogeneous (precise) Markov chain: then we learn something about the transition matrix M
    - an inhomogeneous (precise) Markov chain (without additional assumptions): then we do not learn (almost) anything about the transition matrices M_n
    - a (homogeneous) imprecise Markov chain with the first interpretation (i.e., $M \in \mathcal{M}$): then the estimation of $\mathcal{M}$ does not make sense, but we can use an imprecise Markov chain to describe what we have learned about M

learning from data

- assume that we observe some states $X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n}$ and we consider them as realizations from:
  - a homogeneous (precise) Markov chain: then we learn something about the transition matrix M
  - an inhomogeneous (precise) Markov chain (without additional assumptions): then we do not learn (almost) anything about the transition matrices M_n
  - a (homogeneous) imprecise Markov chain with the first interpretation (i.e., $M \in \mathcal{M}$): then the estimation of $\mathcal{M}$ does not make sense, but we can use an imprecise Markov chain to describe what we have learned about M
  - a (homogeneous) imprecise Markov chain with the second interpretation (i.e., $M_n \in \mathcal{M}$): then the estimation of $\mathcal{M}$ would make sense, but we cannot estimate $\mathcal{M}$ without additional assumptions about the amount of imprecision in $\mathcal{M}$

imprecise inference

- what we learn (from data) about the transition matrix M is described by the (normalized) likelihood function

$$lik(M) = \frac{P_M(X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n})}{\max_{M'} P_{M'}(X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n})}$$

imprecise inference

- what we learn (from data) about the transition matrix M is described by the (normalized) likelihood function

$$lik(M) = \frac{P_M(X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n})}{\max_{M'} P_{M'}(X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n})}$$

- a possibility to obtain an imprecise Markov chain (with the first interpretation, i.e., $M \in \mathcal{M}$) describing what we have learned about M is to choose as set $\mathcal{M}$ of transition matrices the likelihood-based confidence region for M with a certain cutoff point $\beta \in (0, 1)$:

$$\mathcal{M} = \{M : lik(M) > \beta\}$$

imprecise inference

- what we learn (from data) about the transition matrix M is described by the (normalized) likelihood function

$$lik(M) = \frac{P_M(X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n})}{\max_{M'} P_{M'}(X_{i_1} = x_{i_1}, \ldots, X_{i_n} = x_{i_n})}$$

- a possibility to obtain an imprecise Markov chain (with the first interpretation, i.e., $M \in \mathcal{M}$) describing what we have learned about M is to choose as set $\mathcal{M}$ of transition matrices the likelihood-based confidence region for M with a certain cutoff point $\beta \in (0, 1)$:

$$\mathcal{M} = \{M : lik(M) > \beta\}$$

- prior ignorance, learning, and coherence are incompatible: the above idea relaxes coherence during learning (in the sense that the GBR is not satisfied)

imprecise inference

- what we learn (from data) about the transition matrix M is described by the (normalized) likelihood function

$$lik(M) = \frac{P_M(X_{i_1} = x_{i_1}, \dots, X_{i_n} = x_{i_n})}{\max_{M'} P_{M'}(X_{i_1} = x_{i_1}, \dots, X_{i_n} = x_{i_n})}$$

- a possibility to obtain an imprecise Markov chain (with the first interpretation, i.e., $M \in \mathcal{M}$) describing what we have learned about M is to choose as set $\mathcal{M}$ of transition matrices the likelihood-based confidence region for M with a certain cutoff point $\beta \in (0, 1)$:

$$\mathcal{M} = \{M : lik(M) > \beta\}$$

- prior ignorance, learning, and coherence are incompatible: the above idea relaxes coherence during learning (in the sense that the GBR is not satisfied)

- lower and upper previsions corresponding to the imprecise model $\mathcal{M}$ (or more generally, lower and upper bounds on $\mathcal{M}$ for any function of M) can be calculated by a simple algorithm (combining Lagrange multipliers and EM)

example: binary Markov chain ($\mathcal{S} = \{0, 1\}$) with $\beta = 0.15$

- data: $X_1 = 1$
  $P(X_{50} = 1 \mid \text{data}) = [0, 1]$ logit length: ∞
  $P(X_{50} = X_{51} = 1 \mid \text{data}) = [0, 1]$ logit length: ∞

example: binary Markov chain ($\mathcal{S} = \{0, 1\}$) with $\beta = 0.15$

- data: $X_1 = 1$
  $P(X_{50} = 1 \,|\, \text{data}) = [0, 1]$ logit length: ∞
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0, 1]$ logit length: ∞

- data: $X_1 \cdots X_{20} = 11100111100011011110$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.30, 0.85]$ logit length: 2.58
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0.16, 0.75]$ logit length: 2.76

example: binary Markov chain ($\mathcal{S} = \{0, 1\}$) with $\beta = 0.15$

- data: $X_1 = 1$
  $P(X_{50} = 1 \,|\, \text{data}) = [0, 1]$ logit length: ∞
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0, 1]$ logit length: ∞

- data: $X_1 \cdots X_{20} = 11100111100011011110$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.30, 0.85]$ logit length: 2.58
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0.16, 0.75]$ logit length: 2.76

- data: $X_1 \cdots X_{40} = 1110011110001101111010010111011111001001$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.45, 0.76]$ logit length: 1.35
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0.21, 0.61]$ logit length: 1.77

example: binary Markov chain ($\mathcal{S} = \{0, 1\}$) with $\beta = 0.15$

- data: $X_1 = 1$
  $P(X_{50} = 1 \,|\, \text{data}) = [0, 1]$ logit length: ∞
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0, 1]$ logit length: ∞

- data: $X_1 \cdots X_{20} = 11100111100011011110$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.30, 0.85]$ logit length: 2.58
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0.16, 0.75]$ logit length: 2.76

- data: $X_1 \cdots X_{40} = 1110011110001101111010010111011111001001$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.45, 0.76]$ logit length: 1.35
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0.21, 0.61]$ logit length: 1.77

- data (m: MAR):
  $X_1 \cdots X_{40} = 11100111mm001101111m100101110m11110m1001$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.46, 0.80]$ logit length: 1.55

example: binary Markov chain ($\mathcal{S} = \{0, 1\}$) with $\beta = 0.15$

- data: $X_1 = 1$
  $P(X_{50} = 1 \,|\, \text{data}) = [0, 1]$ logit length: ∞
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0, 1]$ logit length: ∞

- data: $X_1 \cdots X_{20} = 11100111100011011110$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.30, 0.85]$ logit length: 2.58
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0.16, 0.75]$ logit length: 2.76

- data: $X_1 \cdots X_{40} = 1110011110001101111010010111011111001001$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.45, 0.76]$ logit length: 1.35
  $P(X_{50} = X_{51} = 1 \,|\, \text{data}) = [0.21, 0.61]$ logit length: 1.77

- data (m: MAR):
  $X_1 \cdots X_{40} = 11100111mm001101111m100101110m11110m1001$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.46, 0.80]$ logit length: 1.55

- data (m: MAR):
  $X_1 \cdots X_{40} = 11mm0111mmm0110m11mm1001m1110m11m1mm1001$
  $P(X_{50} = 1 \,|\, \text{data}) = [0.50, 0.85]$ logit length: 1.73

conclusion

- learning from data with prior (near) ignorance is fundamental for statistical applications of imprecise models

conclusion

- learning from data with prior (near) ignorance is fundamental for statistical applications of imprecise models
- general imprecise approach, easily applied to various statistical models (e.g., continuous-time Markov processes)

conclusion

- learning from data with prior (near) ignorance is fundamental for statistical applications of imprecise models
- general imprecise approach, easily applied to various statistical models (e.g., continuous-time Markov processes)
- very promising algorithm for imprecise inference in discrete (or continuous nonparametric) models

conclusion

- learning from data with prior (near) ignorance is fundamental for statistical applications of imprecise models
- general imprecise approach, easily applied to various statistical models (e.g., continuous-time Markov processes)
- very promising algorithm for imprecise inference in discrete (or continuous nonparametric) models
- can the algorithm be useful for calculating imprecise previsions in general?