Structure learning for CTBN's

Blazej Miasojedow

Institute of Applied Mathematics and Mechanics, University of Warsaw

05 June 2020

[1]

[1] Based on joint works with Wojciech Niemiro (Warsaw/Torun), Wojciech Rejchel (Torun), Maryia Shpak (Lublin)

Outline

1. CTBN

2. Structure learning
   - Full observations
   - Partial observations

Outline

1. CTBN
2. Structure learning
   - Full observations
   - Partial observations

Continuous time Bayesian networks

$X(t)$ multivariate Markov jump process on state $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$ where:

- $(V, \mathcal{E})$ is a directed graph with possible cycles describing dependence structure.
- $\mathcal{X}_v$ space of possible values at node v, assumed to be discrete.

Intensity matrix Q given by conditional intensities

$$Q(x, x\prime) = \begin{cases} Q_v(x_{pa(v)}, x_v, x_v\prime) & \text{if } x_{-v} = x_{-v}\prime \text{ and } x_v \neq x_v\prime \text{ for some } v; \\ 0 & \text{if } x_{-v} \neq x_{-v}\prime \text{ for all } v, \end{cases}$$

where $pa(v)$ denotes the set of parents of node v in the graph $(V, \mathcal{E})$.

Continuous time Bayesian networks

$X(t)$ multivariate Markov jump process on state $\mathcal{X} = \prod_{v \in V} \mathcal{X}_v$ where:

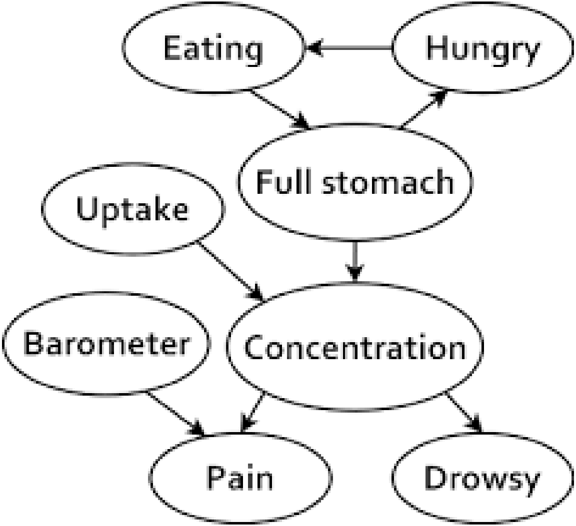
- $(V, \mathcal{E})$ is a directed graph with possible cycles describing dependence structure.
- $\mathcal{X}_v$ space of possible values at node v, assumed to be discrete.

Intensity matrix Q given by conditional intensities

$$Q(x, x\prime) = \begin{cases} Q_v(x_{pa(v)}, x_v, x_v\prime) & \text{if } x_{-v} = x_{-v}\prime \text{ and } x_v \neq x_v\prime \text{ for some } v; \\ 0 & \text{if } x_{-v} \neq x_{-v}\prime \text{ for all } v, \end{cases}$$

where $pa(v)$ denotes the set of parents of node v in the graph $(V, \mathcal{E})$.

Example


Eating
Hungry
Full stomach
Uptake
Barometer
Concentration
Pain
Drowsy

Probability densities of CTBNs

Density can be expressed as a product of conditional densities

$$p(X) = \nu\left(x(0)\right) \prod_{v \in V} p(X_v \| X_{pa(v)}),$$

with

$$p(X_v \| X_{pa(v)}) = \left\{ \prod_{c \in \mathcal{X}_{pa(v)}} \prod_{a \in \mathcal{X}_v} \prod_{\substack{a' \in \mathcal{X}_v \\ a' \neq a}} Q_v(c;\ a, a')^{n_v^T(c;\ a, a')} \right\}$$
$$\left\{ \prod_{c \in \mathcal{X}_{pa(v)}} \prod_{a \in \mathcal{X}_v} \exp\left[-Q_v(c;\ a) t_v^T(c;\ a)\right] \right\},$$

- $n_v^T(c;\ a, a')$ be a number of those jumps from a to a' at node v, which occurred when the parent nodes configuration was c.
- $t_v^T(c;\ a)$ be the length of time when the state of node v was a and the configuration of the parents was c.

1. CTBN
2. Structure learning
   - Full observations
   - Partial observations

Structure learning

Based on observation we want to reconstruct the structure of graph and further estimate conditional intensities matrices. We consider two cases

1. Full trajectory is observed.
2. We observe trajectories only in fixed time points $t_1^{\text{obs}}, \ldots, t_k^{\text{obs}}$ with some noise.

Connections with standard Bayesian networks

- Bayesian networks: consist from inpdependent observations, but graph needs to be acyclic.
- CTBN: dependent observation (Markovian process), no restrictions for graph.
- Easier to formulate thev structure learning problem for CTBNs. No restrictions are required.
- Analysis of methods is more demanding for CTBNs. We need to deal with Markov Jump Processes.

Connections with standard Bayesian networks

- Bayesian networks: consist from inpdependent observations, but graph needs to be acyclic.
- CTBN: dependent observation (Markovian process), no restrictions for graph.
- Easier to formulate thev structure learning problem for CTBNs. No restrictions are required.
- Analysis of methods is more demanding for CTBNs. We need to deal with Markov Jump Processes.

Connections with standard Bayesian networks

- Bayesian networks: consist from inpdependent observations, but graph needs to be acyclic.
- CTBN: dependent observation (Markovian process), no restrictions for graph.
- Easier to formulate thev structure learning problem for CTBNs. No restrictions are required.
- Analysis of methods is more demanding for CTBNs. We need to deal with Markov Jump Processes.

Connections with standard Bayesian networks

- Bayesian networks: consist from inpdependent observations, but graph needs to be acyclic.
- CTBN: dependent observation (Markovian process), no restrictions for graph.
- Easier to formulate thev structure learning problem for CTBNs. No restrictions are required.
- Analysis of methods is more demanding for CTBNs. We need to deal with Markov Jump Processes.

Existing approaches

- Search and score strategy, based on full Bayesian model Nodelman (2007); Acerbi et al. (2014).
- Mean field approximation combined with variational inference Linzner and Koeppl (2018).
- Estimating parameters for full graph in Bayesian setting and removing edges based on marginal posterior probabilities Linzner et al. (2019).

Full observation

Idea:

1. Start with full model.
2. Express
$$\log(Q_v(c,a,a')) = \beta^T Z(c),$$
β is vector of unknown parameter and $Z(c)$ is a vector of dummy variables decoding configuration of all nodes except v.
3. Estimate sparse β by Lasso
$$\arg\min_{\beta} \left\{-\ell(\beta) + \lambda\|\beta\|_1\right\},$$
where ℓ is a likelihood given by
$$\ell(\beta) = \sum_{w\in\mathcal{V}} \sum_{c\in\mathcal{X}_{-w}} \sum_{s\in\mathcal{X}_w} \sum_{\substack{s'\in\mathcal{X}_w \\ s'\neq s}} n_w(c;\ s,s')\beta^w_{s,s'} Z_w(c) - t_w(c;\ s)\exp({\beta^w_{s,s'}}^T Z_w(c)) \tag{1}$$

Example

We consider a binary CTBN with three nodes A, B and C. For the node A we define the function Z_A as

$$Z_A(b,c) = [1, I(b=1), I(c=1)]^\top$$

and β is defined as follows

$$\beta = \left(\beta^A_{0,1}, \beta^A_{1,0}, \beta^B_{0,1}, \beta^B_{1,0}, \beta^C_{0,1}, \beta^C_{1,0}\right)^\top \quad .$$

With slight abuse of notation, the vector $\beta^A_{0,1}$ is given as

$$\beta^A_{0,1} = \left[\beta^A_{0,1}(1), \beta^A_{0,1}(B), \beta^A_{0,1}(C)\right]^\top .$$

Connection between parametrization and structure

In our setting identifying edges in the graph is equivalent to finding non-zero elements of β

$$\beta^{w}_{0,1}(u) \neq 0 \text{ or } \beta^{w}_{1,0}(u) \neq 0 \;\Leftrightarrow\; \text{the edge } u \to w \text{ exists.}$$

Notation and assumptions

- $d_0 = |\mathrm{supp}(\beta)|, \quad S = \mathrm{supp}(\beta), \quad C(\xi) = \{\theta \colon |\theta_{S^C}|_1 \le \xi |\theta_S|_1\}$ for some $\xi > 1$, $\beta_{\min} = \min_k |\beta_k|$
- $$F(\xi) = \inf_{0 \neq \theta \in C(\xi, S)} \sum_{w \in \mathcal{V}} \sum_{s' \neq s} \sum_{c_{S_w} \in \mathcal{X}_{S_w}} \frac{\exp\left(\beta_{s,s'}^{w\top} Z_w(c_{S_w}, 0)\right) \left[\theta_{s,s'}^{w\top} Z_w(c_{S_w}, 0)\right]^2}{|\theta_S|_1 |\theta|_\infty} \qquad (2)$$
- We assume that $F(\xi) > 0$ for some $\xi > 1$
- $\Delta = \max_{s \neq s'} Q(s, s')$

Main result

Theorem 1 (Shpak,Rejchel,BM 2020)

Let $\varepsilon \in (0,1), \xi > 1$ be arbitrary. Suppose that $F(\xi)$ defined in (2) is positive and

$$T > \frac{36\left[(\max_{w\in\mathcal{V}} |S_w| + 1)\log 2 + \log\left(d||\nu||_2/\varepsilon\right)\right]}{\min_{w\in\mathcal{V},s\in\mathcal{X}_w,c_{S_w}\in\mathcal{X}_{S_w}} \pi^2(s,c_{S_w},0)\rho_1}. \tag{3}$$

We also assume that $T\Delta \geq 2$ and

$$2\frac{\xi+1}{\xi-1}\log(K/\varepsilon)\sqrt{\frac{\Delta}{T}} \leq \lambda \leq \frac{2\zeta F(\xi)}{e(\xi+1)|S|}, \tag{4}$$

where $K = 2(2+e^2)d(d-1)$ and $\zeta = \min_{w\in\mathcal{V},s\in\mathcal{X}_w,c_{S_w}\in\mathcal{X}_{S_w}} \pi(s,c_{S_w},0)/2$.

Then with probability at least $1-2\varepsilon$ we have

$$|\hat{\beta} - \beta|_\infty \leq \frac{2e\xi\lambda}{(\xi+1)\zeta F(\xi)}. \tag{5}$$

Consistency of model selection

Corollary 2

Let R denote the right-hand side of the inequality (5). Consider the thresholded Lasso estimator with the set of nonzero coordinates $\hat{S}$. The set $\hat{S}$ contains only those coefficients of the Lasso estimator , which are larger in the absolute value than a pre-specified threshold δ. If $\beta_{min}/2 > \delta \geq R$, then

$$P\left(\hat{S} = S\right) \geq 1 - 2\varepsilon\,.$$

Remarks

- If we forget about constants, Δ and parameters of MJP, i.e. ν, π, ρ_1, ζ etc. in assumptions. Then the estimation error is small, if we have that
$$T \geq \frac{\log^2(d/\varepsilon)|S|^2}{F^2(\xi)}$$
- Conditions (3) and (4) depend also on parameters of MJP. Precisely, they depend on the stationary distribution π and the spectral gap ρ_1, which in general decrease exponentially with d. However, in some specific cases, it can be proved that they decrease polynomially.

CIF vs. F

The cone invertibility factor is defined as

$$\bar{F}(\xi) = \inf_{0 \neq \theta \in C(\xi, S)} \frac{\theta' \nabla^2 \ell(\beta) \theta}{|\theta_S|_1 |\theta|_\infty}.$$

and

$$\theta^T \nabla^2 \ell(\beta) \theta = \frac{1}{T} \sum_{w \in \mathcal{V}} \sum_{c \in \mathcal{X}_{-w}} \sum_{s' \neq s} t_w(c;\ s) \left(\theta_{s,s'}^{w\top} Z_w(c)\right)^2 \exp(\beta_{s,s'}^{w\top} Z_w(c)). \quad (6)$$

- CIF implies "strong convexity" restricted to cone.
- We use classical strategy of proof where positive CIF is required.
- In our case CIF contains a sum of exponentially many r.v.. So we introduce F to overcome this difficulty.

Lower bounds on F

Lemma 3

For every $\xi > 1$ we have with high probablity

$$\bar{F}(\xi) \geq \zeta F(\xi) \geq \frac{\zeta}{\xi A_\beta}\,, \tag{7}$$

where

$$A_\beta = \sum_{w \in \mathcal{V}} \sum_{s' \neq s} \sum_{j : \beta^w_{s,s'}(j) \neq 0} \exp\left(-\beta^w_{s,s'}(j)\right). \tag{8}$$

Sketch of proof

- We use classical technique where it is required to bound $\|\nabla \ell(\beta)\|_\infty$ and $\bar{F}(\xi)$ with high probability.
- To bound $\|\nabla \ell(\beta)\|_\infty$ we derive new concentration inequality for occupation time of MJPs.
- To bound $\bar{F}(\xi)$ we use Lezaud inequality.

Details of implementation

1. Compute lasso estimator on a grid (Estimators for different nodes could be computed in parallel)

$$\hat{\beta}^{w}_{s,s'}(i) = \underset{\theta^{w}_{s,s'}}{\arg\min} \left\{ \ell^{w}_{s,s'}(\theta^{w}_{s,s'}) + \lambda_i |\theta^{w}_{s,s'}|_1 \right\} ,$$

2. Choose λ by BIC:

$$i^* = \underset{1 \le i \le 100}{\arg\min} \left\{ n\ell^{w}_{s,s'}(\hat{\beta}^{w}_{s,s'}(i)) + \log(n) \|\hat{\beta}^{w}_{s,s'}(i)\|_0 \right\} ,$$

3. Choose threshold δ by GIC:

$$\delta^* = \underset{\delta \in \Omega}{\arg\min} \left\{ n\ell^{w}_{s,s'}(\hat{\beta}^{w,\delta}_{s,s'}) + \log(2d(d-1)) \|\hat{\beta}^{w,\delta}_{s,s'}\|_0 \right\} ,$$

Chain example

d	Time	Power	FDR	MD
20	10	0.93	0.21	22.4
	50	0.95	0.07	19.3
50	10	0.86	0.32	61.7
	50	0.88	0.13	49.4

Partial observations

For the partial observation we can analogously define the lasso estimator, but with likelihood of form

$$\ell(\beta) = -\log\left(\int g(y|x)p_\beta(x)\right)dx \ ,$$

where g is distribution of observed y and p_β is density of hidden trajectory of CTBN.

- To solve the lasso problem we can use generalized EM algorithm.
- The expectation step could be done via numerical integration (Nodelman (2007), Linzner and Koeppl (2018),Linzner et al. (2019))
- or by MCMC algorithm Rao and Teh (2012)
- The theoretical analysis of estimator would me much more challenging, because ℓ is no convex anymore.

Partial observations

For the partial observation we can analogously define the lasso estimator, but with likelihood of form

$$\ell(\beta) = -\log\left(\int g(y|x)p_\beta(x)\right)dx\ ,$$

where g is distribution of observed y and p_β is density of hidden trajectory of CTBN.

- To solve the lasso problem we can use generalized EM algorithm.
- The expectation step could be done via numerical integration (Nodelman (2007), Linzner and Koeppl (2018),Linzner et al. (2019))
- or by MCMC algorithm Rao and Teh (2012)
- The theoretical analysis of estimator would me much more challenging, because ℓ is no convex anymore.

Partial observations

For the partial observation we can analogously define the lasso estimator, but with likelihood of form

$$\ell(\beta) = -\log\left(\int g(y|x)p_\beta(x)\right)dx \ ,$$

where g is distribution of observed y and p_β is density of hidden trajectory of CTBN.

- To solve the lasso problem we can use generalized EM algorithm.
- The expectation step could be done via numerical integration (Nodelman (2007), Linzner and Koeppl (2018),Linzner et al. (2019))
- or by MCMC algorithm Rao and Teh (2012)
- The theoretical analysis of estimator would me much more challenging, because ℓ is no convex anymore.

Partial observations

For the partial observation we can analogously define the lasso estimator, but with likelihood of form

$$\ell(\beta) = -\log \left(\int g(y|x) p_\beta(x) \right) dx \ ,$$

where g is distribution of observed y and p_β is density of hidden trajectory of CTBN.

- To solve the lasso problem we can use generalized EM algorithm.
- The expectation step could be done via numerical integration (Nodelman (2007), Linzner and Koeppl (2018),Linzner et al. (2019))
- or by MCMC algorithm Rao and Teh (2012)
- The theoretical analysis of estimator would me much more challenging, because ℓ is no convex anymore.

Thank you!