Sparse Markov Models for High-Dimensional Inference

Guilherme Ost

Federal University of Rio de Janeiro

CYU ECODEP

Jun, 2022

Joint work with

D.Y. Takahashi
(Brain Institute/UFRN)

Data: n observations $X_1, \ldots, X_n$ of a stationary binary Markov chain of order d.

Data: n observations $X_1, \ldots, X_n$ of a stationary binary Markov chain of order d.

Transition probabilities: $p(x) = \mathbb{P}(X_t = 1 | X_{t-d:t-1} = x)$ for $x \in \{0, 1\}^d$.

Data: n observations $X_1, \ldots, X_n$ of a stationary binary Markov chain of order d.

Transition probabilities: $p(x) = \mathbb{P}(X_t = 1 | X_{t-d:t-1} = x)$ for $x \in \{0, 1\}^d$.

Classical question: how to estimate the transition probabilities $p(x)$ given the data?

Data: n observations $X_1, \ldots, X_n$ of a stationary binary Markov chain of order d.

Transition probabilities: $p(x) = \mathbb{P}(X_t = 1 | X_{t-d:t-1} = x)$ for $x \in \{0,1\}^d$.

Classical question: how to estimate the transition probabilities $p(x)$ given the data?

The MLE of $p(x)$ computed from the data is given by

$$\hat{p}_n(x) = \frac{N_n(x,1)}{N_n(x,0) + N_n(x,1)} = \frac{N_n(x,1)}{\bar{N}_n(x)},$$

where $N_n(x,b) = |\{d+1 \leq t \leq n : X_{t-d:t-1} = x, X_t = b\}|$.

Data: n observations $X_1, \ldots, X_n$ of a stationary binary Markov chain of order d.

Transition probabilities: $p(x) = \mathbb{P}(X_t = 1 | X_{t-d:t-1} = x)$ for $x \in \{0, 1\}^d$.

Classical question: how to estimate the transition probabilities $p(x)$ given the data?

The MLE of $p(x)$ computed from the data is given by

$$\hat{p}_n(x) = \frac{N_n(x, 1)}{N_n(x, 0) + N_n(x, 1)} = \frac{N_n(x, 1)}{\bar{N}_n(x)},$$

where $N_n(x, b) = |\{d + 1 \leq t \leq n : X_{t-d:t-1} = x, X_t = b\}|$.

Focus on the high-dimensional setting: $d = d_n$ and $p(x) = p_n(x)$.

In this full generality, $p(x)$ can be estimated only if $d \leq C \log_2 n$.

In this full generality, $p(x)$ can be estimated only if $d \leq C \log_2 n$.

There are two related ways to convince of this:

1. The number of free parameters $Dim_{MC}(d) = 2^d$ grows exponentially with d.

In this full generality, $p(x)$ can be estimated only if $d \leq C \log_2 n$.

There are two related ways to convince of this:

1. The number of free parameters $Dim_{MC}(d) = 2^d$ grows exponentially with d.

2. For $\hat{p}_n(x)$ to have any meaning, we need that $\bar{N}_n(x) \geq 1$.

In this full generality, $p(x)$ can be estimated only if $d \leq C \log_2 n$.

There are two related ways to convince of this:

1. The number of free parameters $Dim_{MC}(d) = 2^d$ grows exponentially with d.

2. For $\hat{p}_n(x)$ to have any meaning, we need that $\bar{N}_n(x) \geq 1$. By ergodicity,

$$\bar{N}_n(x) \approx n\mathbb{P}(X_{1:d} = x).$$

In this full generality, $p(x)$ can be estimated only if $d \leq C \log_2 n$.

There are two related ways to convince of this:

1. The number of free parameters $Dim_{MC}(d) = 2^d$ grows exponentially with d.

2. For $\hat{p}_n(x)$ to have any meaning, we need that $\bar{N}_n(x) \geq 1$. By ergodicity,

$$\bar{N}_n(x) \approx n\mathbb{P}(X_{1:d} = x).$$

If the transition probabilities are bounded below from zero, then $\exists\ c > 0$ such that

$$\mathbb{P}(X_{1:d} = x) < e^{-cd}.$$

In this full generality, $p(x)$ can be estimated only if $d \leq C \log_2 n$.

There are two related ways to convince of this:

1. The number of free parameters $Dim_{MC}(d) = 2^d$ grows exponentially with d.

2. For $\hat{p}_n(x)$ to have any meaning, we need that $\bar{N}_n(x) \geq 1$. By ergodicity,

$$\bar{N}_n(x) \approx n\mathbb{P}(X_{1:d} = x).$$

If the transition probabilities are bounded below from zero, then $\exists\ c > 0$ such that

$$\mathbb{P}(X_{1:d} = x) < e^{-cd}.$$

Hence, we need $1 \leq ne^{-cd}$ implying that $d \leq C \log_2 n$.

In this full generality, $p(x)$ can be estimated only if $d \leq C \log_2 n$.

There are two related ways to convince of this:

1. The number of free parameters $Dim_{MC}(d) = 2^d$ grows exponentially with d.

2. For $\hat{p}_n(x)$ to have any meaning, we need that $\bar{N}_n(x) \geq 1$. By ergodicity,

$$\bar{N}_n(x) \approx n\mathbb{P}(X_{1:d} = x).$$

If the transition probabilities are bounded below from zero, then $\exists\ c > 0$ such that

$$\mathbb{P}(X_{1:d} = x) < e^{-cd}.$$

Hence, we need $1 \leq ne^{-cd}$ implying that $d \leq C \log_2 n$. Need to seek for sparse Markov chains!

Two examples of sparse $\{0,1\}$-valued Markov chains

Minimal Markov Models (MMM) are Markov chains of order d such that there exist a partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of $\{0,1\}^d$ with the property that

$$p(x) = p(y) \text{ if and only if } x, y \in \mathcal{C}_i.$$

Two examples of sparse $\{0,1\}$-valued Markov chains

Minimal Markov Models (MMM) are Markov chains of order d such that there exist a partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of $\{0,1\}^d$ with the property that

$$p(x) = p(y) \text{ if and only if } x, y \in \mathcal{C}_i.$$

The dimension of a MMM is $Dim_{MMM}(d) = K$. Sparse when $K \ll 2^d$.

Two examples of sparse $\{0, 1\}$-valued Markov chains

Minimal Markov Models (MMM) are Markov chains of order d such that there exist a partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of $\{0, 1\}^d$ with the property that

$$p(x) = p(y) \text{ if and only if } x, y \in \mathcal{C}_i.$$

The dimension of a MMM is $Dim_{MMM}(d) = K$. Sparse when $K \ll 2^d$.

Variable length Markov chains (VLMC) are MMM for which the partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$ is "given by a irreducible tree".

Two examples of sparse $\{0,1\}$-valued Markov chains

Minimal Markov Models (MMM) are Markov chains of order d such that there exist a partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$ of $\{0,1\}^d$ with the property that

$$p(x) = p(y) \text{ if and only if } x, y \in \mathcal{C}_i.$$

The dimension of a MMM is $Dim_{MMM}(d) = K$. Sparse when $K \ll 2^d$.

Variable length Markov chains (VLMC) are MMM for which the partition $\mathcal{C}_1, \ldots, \mathcal{C}_K$ is "given by a irreducible tree".

Without further hypothesis, $p(x)$ can be estimated still only if $d \leq C \log_2 n$ (by the point 2 above.)

What if $d \gg C \log_2 n$?

What if $d \gg C \log_2 n$?

Why this regime is important? Many natural phenomena have very long memory!

What if $d \gg C \log_2 n$?

Why this regime is important? Many natural phenomena have very long memory!

In this talk: we suppose $d = \beta n$ with $\beta \in (0, 1)$ and focus on another class of sparse Markov chains, called Mixture Transition Distribution (MTD) models.

MTD models have been introduced by A. Raftery ('85). For applications see A. Berchtold & Raftery ('02).

MTD models

Markov chains of order d such that

$$p(x) = \lambda_0 p_0 + \sum_{j=-d}^{-1} \lambda_j p_j(x_j)$$

where: $x = (x_{-d}, \ldots, x_{-1})$ and

- $0 \leq p_0, p_j(a) \leq 1$ for all $j \in \{-d, \ldots, -1\}$ and $a \in \{0, 1\}$.
- $\lambda_0, \lambda_1, \ldots, \lambda_{-d} \in [0, 1]$ such that $\sum_{j=-d}^{0} \lambda_j = 1$.

MTD models

Markov chains of order d such that

$$p(x) = \lambda_0 p_0 + \sum_{j=-d}^{-1} \lambda_j p_j(x_j)$$

where: $x = (x_{-d}, \ldots, x_{-1})$ and

- $0 \leq p_0, p_j(a) \leq 1$ for all $j \in \{-d, \ldots, -1\}$ and $a \in \{0, 1\}$.
- $\lambda_0, \lambda_1, \ldots, \lambda_{-d} \in [0, 1]$ such that $\sum_{j=-d}^{0} \lambda_j = 1$.

For each lag $j \in \{-d, \ldots, -1\}$, let $\delta_j = \lambda_j |p_j(1) - p_j(0)|$.

MTD models

Markov chains of order d such that

$$p(x) = \lambda_0 p_0 + \sum_{j=-d}^{-1} \lambda_j p_j(x_j)$$

where: $x = (x_{-d}, \ldots, x_{-1})$ and

- $0 \leq p_0, p_j(a) \leq 1$ for all $j \in \{-d, \ldots, -1\}$ and $a \in \{0, 1\}$.
- $\lambda_0, \lambda_1, \ldots, \lambda_{-d} \in [0, 1]$ such that $\sum_{j=-d}^{0} \lambda_j = 1$.

For each lag $j \in \{-d, \ldots, -1\}$, let $\delta_j = \lambda_j |p_j(1) - p_j(0)|$.

Denote $\Lambda = \{j \in \{-d, \ldots, -1\} : \delta_j > 0\}$ (set of relevant lags).

MTD models

Markov chains of order d such that

$$p(x) = \lambda_0 p_0 + \sum_{j=-d}^{-1} \lambda_j p_j(x_j)$$

where: $x = (x_{-d}, \ldots, x_{-1})$ and

- $0 \leq p_0, p_j(a) \leq 1$ for all $j \in \{-d, \ldots, -1\}$ and $a \in \{0, 1\}$.
- $\lambda_0, \lambda_1, \ldots, \lambda_{-d} \in [0, 1]$ such that $\sum_{j=-d}^{0} \lambda_j = 1$.

For each lag $j \in \{-d, \ldots, -1\}$, let $\delta_j = \lambda_j |p_j(1) - p_j(0)|$.

Denote $\Lambda = \{j \in \{-d, \ldots, -1\} : \delta_j > 0\}$ (set of relevant lags).

Note that $p(x) = p(x_\Lambda)$ and $Dim_{MTD}(d) = 3|\Lambda| + 1$.

Idea to estimate $p(x)$ for MTD

First, estimate Λ from the data. Denote $\hat{\Lambda}_n$ an estimator of Λ.

Idea to estimate $p(x)$ for MTD

First, estimate Λ from the data. Denote $\hat{\Lambda}_n$ an estimator of Λ.

Then, compute $\hat{p}_n(x_{\hat{\Lambda}_n})$.

Idea to estimate $p(x)$ for MTD

First, estimate Λ from the data. Denote $\hat{\Lambda}_n$ an estimator of Λ.

Then, compute $\hat{p}_n(x_{\hat{\Lambda}_n})$.

Statistical lag selection: how to estimate efficiently Λ from the data?

Idea to estimate $p(x)$ for MTD

First, estimate Λ from the data. Denote $\hat{\Lambda}_n$ an estimator of Λ.

Then, compute $\hat{p}_n(x_{\hat{\Lambda}_n})$.

Statistical lag selection: how to estimate efficiently Λ from the data?

Remark: the behavior of $\min_{j\in\Lambda}\delta_j^2$ measures how difficult is to estimate Λ.

Idea to estimate $p(x)$ for MTD

First, estimate Λ from the data. Denote $\hat{\Lambda}_n$ an estimator of Λ.

Then, compute $\hat{p}_n(x_{\hat{\Lambda}_n})$.

Statistical lag selection: how to estimate efficiently Λ from the data?

Remark: the behavior of $\min_{j\in\Lambda} \delta_j^2$ measures how difficult is to estimate Λ.

Indeed, lag selection is possible (in the minimax sense) only if

$$\min_{j\in\Lambda} \delta_j^2 \geq C\frac{\log(n)}{n}.$$

Goal of this talk:

- to present an efficient estimator of the set of relevant lags Λ, based on a sample $X_{1:n}$ of a MTD model with order d.
- to provide some theoretical guarantees in the high-dimensional regime $\Lambda = \Lambda_n$ and $d = d_n = \beta n$ for some $\beta \in (0, 1)$.

Goal of this talk:

- to present an efficient estimator of the set of relevant lags Λ, based on a sample $X_{1:n}$ of a MTD model with order d.
- to provide some theoretical guarantees in the high-dimensional regime $\Lambda = \Lambda_n$ and $d = d_n = \beta n$ for some $\beta \in (0, 1)$.

To estimate Λ, we propose to use the *Forward Stepwise and Cut* (FSC) estimator.

Goal of this talk:

- to present an efficient estimator of the set of relevant lags Λ, based on a sample $X_{1:n}$ of a MTD model with order d.
- to provide some theoretical guarantees in the high-dimensional regime $\Lambda = \Lambda_n$ and $d = d_n = \beta n$ for some $\beta \in (0, 1)$.

To estimate Λ, we propose to use the *Forward Stepwise and Cut* (FSC) estimator.

For a sample $X_{1:n}$, integer $m < n$, $S \subseteq \{-d, \ldots, -1\}$ and $x_S \in \{0, 1\}^S$, let

$$\hat{p}_{m,n}(x_S) = \begin{cases} \frac{N_{m,n}(x_S,1)}{\bar{N}_{m,n}(x_S)}, & \text{if } \bar{N}_{m,n}(x_S) > 0, \\ 1/2, & \text{otherwise} \end{cases},$$

In the definition of $\hat{p}_{m,n}(x_S)$ the countings are over $X_{m+1:n}$.

FSC estimator

The FSC estimator is defined as follows.

Step 1 (FS). From $X_{1:m}$, build a random set $\hat{S}_m$ such that $\Lambda \subseteq \hat{S}_m$ with high probability.

FSC estimator

The FSC estimator is defined as follows.

Step 1 (FS). From $X_{1:m}$, build a random set $\hat{S}_m$ such that $\Lambda \subseteq \hat{S}_m$ with high probability.

Step 2 (CUT). For each $j \in \hat{S}_m$, remove j from $\hat{S}_m$ only if

$$|\hat{p}_{m,n}(x_{\hat{S}_m}) - \hat{p}_{m,n}(y_{\hat{S}_m})| < t_{m,n}(x_{\hat{S}_m}, y_{\hat{S}_m}),$$

for all $x_{\hat{S}_m}, y_{\hat{S}_m} \in A^{\hat{S}_m}$ $s.t.$ $x_k = y_k$ for all $k \in \hat{S}_m \setminus \{j\}$.

FSC estimator

The FSC estimator is defined as follows.

Step 1 (FS). From $X_{1:m}$, build a random set $\hat{S}_m$ such that $\Lambda \subseteq \hat{S}_m$ with high probability.

Step 2 (CUT). For each $j \in \hat{S}_m$, remove j from $\hat{S}_m$ only if

$$|\hat{p}_{m,n}(x_{\hat{S}_m}) - \hat{p}_{m,n}(y_{\hat{S}_m})| < t_{m,n}(x_{\hat{S}_m}, y_{\hat{S}_m}),$$

for all $x_{\hat{S}_m}, y_{\hat{S}_m} \in A^{\hat{S}_m}$ $s.t.$ $x_k = y_k$ for all $k \in \hat{S}_m \setminus \{j\}$.

Output $\hat{\Lambda}_n =$ All lags not removed in the CUT step.

Choice of the random threshold

For $S \subseteq \{-d, \ldots, -1\}$, $x_S \in \{0,1\}^S$, we take $t_{m,n}(x_S, y_S) = s_{m,n}(x_S) + s_{m,n}(y_S)$, where

$$s_{m,n}(x_S) = \sqrt{\frac{2\alpha(1+\varepsilon)V_{m,n}(x_S)}{\bar{N}_{m,n}(x_S)}} + \frac{2\alpha}{3\bar{N}_{m,n}(x_S)},$$

with $\alpha, \varepsilon > 0$, $\mu \in (0,3)$ s.t. $\mu > \psi(\mu) = e^{\mu} - 1 - \mu$ and

$$V_{m,n}(x_S) = \frac{\mu}{\mu - \psi(\mu)}\hat{p}_{m,n}(x_S) + \frac{\alpha}{\bar{N}_{m,n}(x_S)(\mu - \psi(\mu))}.$$

Choice of the random threshold

For $S \subseteq \{-d, \dots, -1\}$, $x_S \in \{0,1\}^S$, we take $t_{m,n}(x_S, y_S) = s_{m,n}(x_S) + s_{m,n}(y_S)$, where

$$s_{m,n}(x_S) = \sqrt{\frac{2\alpha(1+\varepsilon)V_{m,n}(x_S)}{\bar{N}_{m,n}(x_S)}} + \frac{2\alpha}{3\bar{N}_{m,n}(x_S)},$$

with $\alpha, \varepsilon > 0$, $\mu \in (0,3)$ s.t. $\mu > \psi(\mu) = e^{\mu} - 1 - \mu$ and

$$V_{m,n}(x_S) = \frac{\mu}{\mu - \psi(\mu)} \hat{p}_{m,n}(x_S) + \frac{\alpha}{\bar{N}_{m,n}(x_S)(\mu - \psi(\mu))}.$$

The choice of $s_{m,n}(x_S)$ is based on a Martingale concentration inequality.

How do we build $\hat{S}_m$?

For $S \subseteq \{-d, \ldots, -1\}$ and $j \notin S$, let $\bar{\nu}_{j,S} = \mathbb{E}\left[|Cov_{X_S}(X_0, X_j)|\right].$

How do we build $\hat{S}_m$?

For $S \subseteq \{-d, \ldots, -1\}$ and $j \notin S$, let $\bar{\nu}_{j,S} = \mathbb{E}\left[\left| Cov_{X_S}(X_0, X_j) \right|\right]$.

Notice that $\max_{j \in S^c} \bar{\nu}_{j,S} = 0$ if $\Lambda \subseteq S$.

How do we build $\hat{S}_m$?

For $S \subseteq \{-d, \ldots, -1\}$ and $j \notin S$, let $\bar{\nu}_{j,S} = \mathbb{E}\left[|Cov_{X_S}(X_0, X_j)|\right]$.

Notice that $\max_{j \in S^c} \bar{\nu}_{j,S} = 0$ if $\Lambda \subseteq S$.

Assumption 1. $\mathbb{P}(X_S = x_S) > 0$ for all $S \subseteq \{-d, \ldots, -1\}$ and $x_S \in \{0, 1\}^S$.

How do we build $\hat{S}_m$?

For $S \subseteq \{-d, \dots, -1\}$ and $j \notin S$, let $\bar{\nu}_{j,S} = \mathbb{E}\left[|Cov_{X_S}(X_0, X_j)|\right]$.

Notice that $\max_{j \in S^c} \bar{\nu}_{j,S} = 0$ if $\Lambda \subseteq S$.

Assumption 1. $\mathbb{P}(X_S = x_S) > 0$ for all $S \subseteq \{-d, \dots, -1\}$ and $x_S \in \{0,1\}^S$.

Proposition 1. Under Assumption 1 there exists $\kappa > 0$ such that the following property holds: for all $S \subseteq \{-d, \dots, -1\}$ with $\Lambda \not\subseteq S$, it holds that

$$\max_{j \in S^c} \bar{\nu}_{j,S} \geq \max_{j \in \Lambda \setminus S} \bar{\nu}_{j,S} \geq \kappa$$

Denote $\hat{\nu}_{m,j,S}$ the empirical estimate of $\bar{\nu}_{j,S}$ computed from $X_{1:m}$.

Denote $\hat{\nu}_{m,j,S}$ the empirical estimate of $\bar{\nu}_{j,S}$ computed from $X_{1:m}$.

To build $\hat{S}_m$, we do as follows. Fix $0 \leq \ell \leq d$.

1. Set $\hat{S}_m = \emptyset$.
2. While $|\hat{S}_m| < \ell$, compute $j \in \arg\max_{k \in \hat{S}_m^c} \hat{\nu}_{m,k,\hat{S}_m}$ and include j in $\hat{S}_m$.

Theoretical guarantees of FSC estimator.

Theorem. Take $m = n/2$ and assume $d = \beta m$ for $\beta \in (0, 1)$ and suppose $\lambda_0 > 0$, $0 < p_0 < 1$ and that the following conditions hold:

Theoretical guarantees of FSC estimator.

Theorem. Take $m = n/2$ and assume $d = \beta m$ for $\beta \in (0,1)$ and suppose $\lambda_0 > 0$, $0 < p_0 < 1$ and that the following conditions hold:

- $\exists\ \Gamma_1 \in (0,1]$ s.t. for all $S \subset \{-d,\dots,-1\}$ such that $\Lambda \not\subseteq S$ and $k \in \Lambda \setminus S$,

$$\max_{x_S \in \{0,1\}^S} \sum_{j \in \Lambda \setminus S \cup \{k\}} \frac{\delta_j}{\delta_k} |\mathbb{P}_{x_S}(X_j = 1 | X_k = 0) - \mathbb{P}_{x_S}(X_j = 1 | X_k = 1)| \leq (1 - \Gamma_1).$$

Theoretical guarantees of FSC estimator.

Theorem. Take $m = n/2$ and assume $d = \beta m$ for $\beta \in (0,1)$ and suppose $\lambda_0 > 0$, $0 < p_0 < 1$ and that the following conditions hold:

- $\exists\ \Gamma_1 \in (0,1]$ s.t. for all $S \subset \{-d, \ldots, -1\}$ such that $\Lambda \not\subseteq S$ and $k \in \Lambda \setminus S$,

$$\max_{x_S \in \{0,1\}^S} \sum_{j \in \Lambda \setminus S \cup \{k\}} \frac{\delta_j}{\delta_k} |\mathbb{P}_{x_S}(X_j = 1 | X_k = 0) - \mathbb{P}_{x_S}(X_j = 1 | X_k = 1)| \leq (1 - \Gamma_1).$$

- $\exists\ \Gamma_2 \in (0,1]$ s.t. for all $S \subset \{-d, \ldots, -1\}$ such that $\Lambda \subset S$ and $k \notin \Lambda$,

$$\sum_{j \in \Lambda \setminus S} \max_{x_S \in \{0,1\}^S} |\mathbb{P}_{x_S}(X_k = 1 | X_j = 0) - \mathbb{P}_{x_S}(X_k = 1 | X_j = 1)| \leq \Gamma_2.$$

Suppose also $|\Lambda| \leq L$ with L known and let $\hat{\Lambda}_n$ be the FSC estimator constructed with parameters $\ell = L$ and $\alpha = (1 + \eta)\log(n)$ for $\eta > 0$.

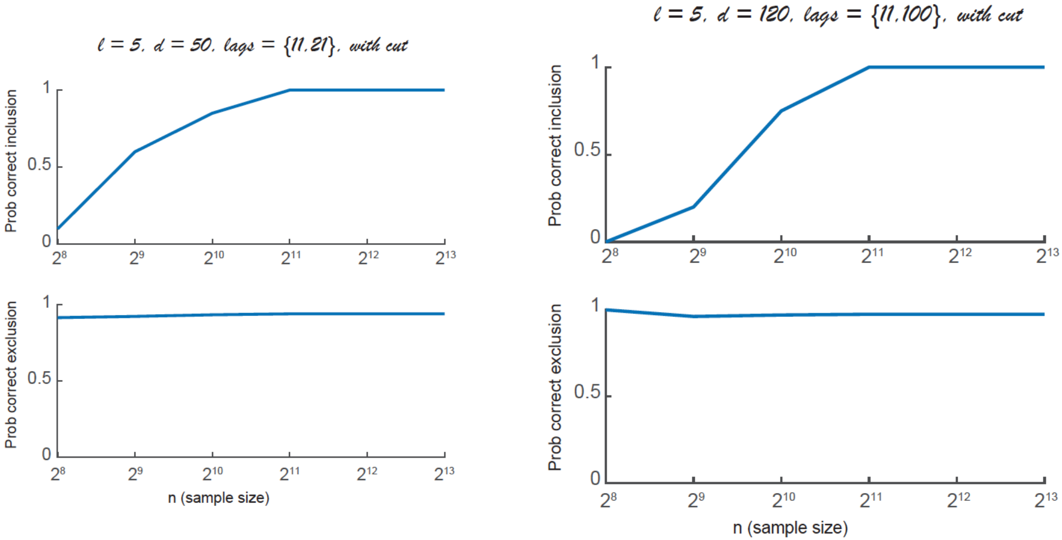
Theoretical guarantees of FSC estimator.

Theorem. Take $m = n/2$ and assume $d = \beta m$ for $\beta \in (0, 1)$ and suppose $\lambda_0 > 0$, $0 < p_0 < 1$ and that the following conditions hold:

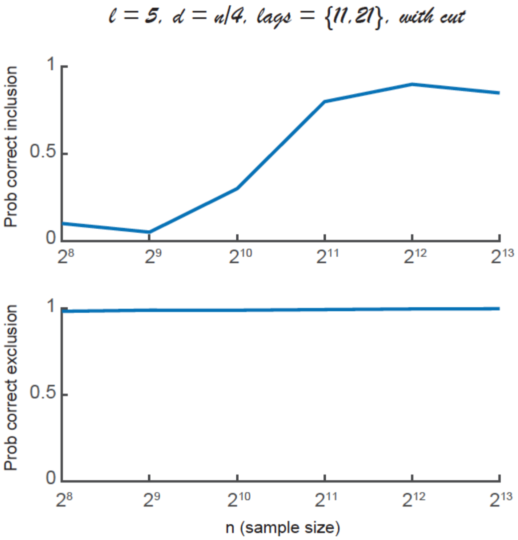
- $\exists\ \Gamma_1 \in (0, 1]$ s.t. for all $S \subset \{-d, \ldots, -1\}$ such that $\Lambda \not\subseteq S$ and $k \in \Lambda \setminus S$,
$$\max_{x_S \in \{0,1\}^S} \sum_{j \in \Lambda \setminus S \cup \{k\}} \frac{\delta_j}{\delta_k} |\mathbb{P}_{x_S}(X_j = 1 | X_k = 0) - \mathbb{P}_{x_S}(X_j = 1 | X_k = 1)| \leq (1 - \Gamma_1).$$
- $\exists\ \Gamma_2 \in (0, 1]$ s.t. for all $S \subset \{-d, \ldots, -1\}$ such that $\Lambda \subset S$ and $k \notin \Lambda$,
$$\sum_{j \in \Lambda \setminus S} \max_{x_S \in \{0,1\}^S} |\mathbb{P}_{x_S}(X_k = 1 | X_j = 0) - \mathbb{P}_{x_S}(X_k = 1 | X_j = 1)| \leq \Gamma_2.$$

Suppose also $|\Lambda| \leq L$ with L known and let $\hat{\Lambda}_n$ be the FSC estimator constructed with parameters $\ell = L$ and $\alpha = (1 + \eta) \log(n)$ for $\eta > 0$. Then $\exists$ a constant $C > 0$ such that
$$\mathbb{P}(\hat{\Lambda}_n \neq \Lambda) \to 0 \text{ as } n \to \infty,$$
as long as
$$\min_{j \in \Lambda} \delta_j^2 \geq C \frac{\log(n)}{n}.$$

Simulations: FSC estimator



Simulations: FSC estimator



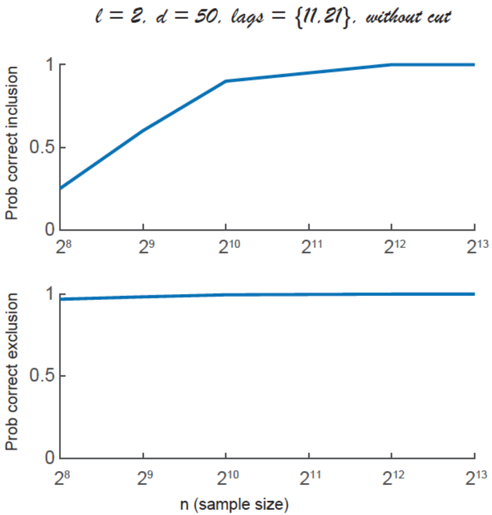
Simulations: transition probability estimation

MTD model used: $p(x) = \lambda_0 p_0 + \lambda_i p_i(x_i) + \lambda_j p_j(x_j)$ where $\lambda_0 = 0.2$, $p_0 = 0.5$, $\lambda_i = \lambda_j = 0.4$, $1 - p_i(0) = p_i(1) = 1 - p_j(0) = p_j(1) = 0.7$.

For each choice of i, j, d, and n we simulated 100 realizations. For each realization, we estimated the transition probability $p(0|0^d)$.

Model parameter			Method	Sample size (n)					
i	j	d		**256**	**512**	**1024**	**2048**	**4096**	**8192**
1	5	5	FSC(2)	0.0774	0.0682	0.0506	0.0286	0.0174	0.0133
1	5	5	FSC(5)	0.0745	0.0835	0.0602	0.0426	0.0222	0.0129
1	5	5	PCP	0.0965	0.0786	0.0577	0.0432	0.0242	0.0131
1	5	5	Naive	0.1518	0.0933	0.0624	0.0455	0.0340	0.0252
1	5	10	FSC(5)	0.0836	0.0842	0.0659	0.0425	0.0228	0.0141
1	10	15	FSC(5)	0.0864	0.0781	0.0641	0.0438	0.0249	0.0151
1	15	20	FSC(5)	0.0682	0.0802	0.0778	0.0534	0.0285	0.0138
11	100	120	FSC(5)	-	-	0.0838	0.0647	0.0312	0.0169
1	10	n/8	FSC(5)	0.0563	0.0543	0.0780	0.0698	0.0504	0.0105

Simulations: FSC without CUT



Final comments

We could estimate Λ by

$$\hat{\Lambda}_{BIC} = \arg \min_{S \in \mathcal{P}(\{-d,\ldots,-1\})} \left\{ -\log ML_S(X_1, \ldots, X_n) + \frac{(3|\Lambda| + 1)}{2} \log(n) \right\}.$$

Can we compute $\hat{\Lambda}_{BIC}$ efficiently? The models are not nested!

Final comments

We could estimate Λ by

$$\hat{\Lambda}_{BIC} = \arg \min_{S \in \mathcal{P}(\{-d,\dots,-1\})} \left\{ -\log ML_S(X_1, \dots, X_n) + \frac{(3|\Lambda| + 1)}{2} \log(n) \right\}.$$

Can we compute $\hat{\Lambda}_{BIC}$ efficiently? The models are not nested!

What about multivariate MTD models?