Sparse Markov Models for High-Dimensional Inference

Guilherme Ost

Federal University of Rio de Janeiro

Probability Webinar - IM - UFRJ

March, 2022

Joint work with

D.Y. Takahashi
(Brain Institute/UFRN)

Let A be a finite subset of $\mathbb{R}$. (alphabet).

Let A be a finite subset of $\mathbb{R}$. (alphabet).

For $k \geq 1$, denote $x_{-k:-1} = (x_{-k}, \ldots, x_{-1}) \in A^{\{-k,\ldots,-1\}}$.

Let A be a finite subset of $\mathbb{R}$. (alphabet).

For $k \geq 1$, denote $x_{-k:-1} = (x_{-k}, \ldots, x_{-1}) \in A^{\{-k,\ldots,-1\}}$.

Consider a stationary Markov chain $(X_t)_{t\in\mathbb{Z}}$ of order $d \geq 1$:

$$\mathbb{P}(X_t = a | X_{t-k:t-1} = x_{-k:-1}) = \mathbb{P}(X_t = a | X_{t-d:t-1} = x_{-d:-1}) \ \forall \ t \in \mathbb{Z}, \ k > d,$$

$a \in A$ and $x_{-k:-1} \in A^{\{-k,\ldots,-1\}}$ such that $\mathbb{P}(X_{t-k:t-1} = x_{t-k:t-1}) > 0$.

Let A be a finite subset of $\mathbb{R}$. (alphabet).

For $k \geq 1$, denote $x_{-k:-1} = (x_{-k}, \ldots, x_{-1}) \in A^{\{-k,\ldots,-1\}}$.

Consider a stationary Markov chain $(X_t)_{t\in\mathbb{Z}}$ of order $d \geq 1$:

$$\mathbb{P}(X_t = a | X_{t-k:t-1} = x_{-k:-1}) = \mathbb{P}(X_t = a | X_{t-d:t-1} = x_{-d:-1}) \ \forall \ t \in \mathbb{Z}, \ k > d,$$

$a \in A$ and $x_{-k:-1} \in A^{\{-k,\ldots,-1\}}$ such that $\mathbb{P}(X_{t-k:t-1} = x_{t-k:t-1}) > 0$.

Denote $p(a|x_{-d:-1}) = \mathbb{P}(X_0 = a | X_{-d:-1} = x_{-d:-1})$ (transition probabilities).

Classical statistical questions: given a sample $X_{1:n}$ of a Markov chain,

- How to estimate the order d?
- How to estimate the transition probabilities $p(a|x_{-d:-1})$?
- Can we provide confidence intervals for $p(a|x_{-d:-1})$?

Classical statistical questions: given a sample $X_{1:n}$ of a Markov chain,

- How to estimate the order d?
- How to estimate the transition probabilities $p(a|x_{-d:-1})$?
- Can we provide confidence intervals for $p(a|x_{-d:-1})$?

Focus on the high-dimensional setting: $d = d_n$ and $p(a|x_{-d:-1}) = p_n(a|x_{-d_n:-1})$.

Classical statistical questions: given a sample $X_{1:n}$ of a Markov chain,

- How to estimate the order d?
- How to estimate the transition probabilities $p(a|x_{-d:-1})$?
- Can we provide confidence intervals for $p(a|x_{-d:-1})$?

Focus on the high-dimensional setting: $d = d_n$ and $p(a|x_{-d:-1}) = p_n(a|x_{-d_n:-1})$.

One challenge is the curse of dimensionality: $Dim_{MC}(d) = |A|^d(|A| - 1)$ grows exponentially with d.

Classical statistical questions: given a sample $X_{1:n}$ of a Markov chain,

- How to estimate the order d?
- How to estimate the transition probabilities $p(a|x_{-d:-1})$?
- Can we provide confidence intervals for $p(a|x_{-d:-1})$?

Focus on the high-dimensional setting: $d = d_n$ and $p(a|x_{-d:-1}) = p_n(a|x_{-d_n:-1})$.

One challenge is the curse of dimensionality: $Dim_{MC}(d) = |A|^d(|A| - 1)$ grows exponentially with d.

Typically, in the high-dimensional setting $Dim_{MC}(d_n) \gg n$. Need to seek for low dimensional (sparse) Markov chains!

Two examples of sparse Markov chains

Variable length Markov chains (VLMC) are Markov chains of order d such that

$$p(a|x_{-d:-1}) = p(a|x_{-\ell:-1}) \text{ for some } \ell = \ell(x_{-d:-1}).$$

Denote $\tau = \{x_{-\ell:-1} : \ell = \ell(x_{-d:-1}), x_{-d:-1} \in A^{\{-d,\dots,-1\}}\}$.

Two examples of sparse Markov chains

Variable length Markov chains (VLMC) are Markov chains of order d such that

$$p(a|x_{-d:-1}) = p(a|x_{-\ell:-1}) \text{ for some } \ell = \ell(x_{-d:-1}).$$

Denote $\tau = \{x_{-\ell:-1} : \ell = \ell(x_{-d:-1}), x_{-d:-1} \in A^{\{-d,\dots,-1\}}\}$.

The dimension of an VLMC is $Dim_{VLMC}(d) = |\tau|(|A|-1)$. Typically, $|\tau| \ll |A|^d$.

Two examples of sparse Markov chains

Variable length Markov chains (VLMC) are Markov chains of order d such that

$$p(a|x_{-d:-1}) = p(a|x_{-\ell:-1}) \text{ for some } \ell = \ell(x_{-d:-1}).$$

Denote $\tau = \{x_{-\ell:-1} : \ell = \ell(x_{-d:-1}), x_{-d:-1} \in A^{\{-d,\dots,-1\}}\}$.

The dimension of an VLMC is $Dim_{VLMC}(d) = |\tau|(|A| - 1)$. Typically, $|\tau| \ll |A|^d$.

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

$$p(a|x_{-d:-1}) = p(a|y_{-d:-1}) \text{ if and only if } x_{-d:-1}, y_{-d:-1} \in \mathcal{C}_i.$$

Two examples of sparse Markov chains

Variable length Markov chains (VLMC) are Markov chains of order d such that

$$p(a|x_{-d:-1}) = p(a|x_{-\ell:-1}) \text{ for some } \ell = \ell(x_{-d:-1}).$$

Denote $\tau = \{x_{-\ell:-1} : \ell = \ell(x_{-d:-1}), x_{-d:-1} \in A^{\{-d,\dots,-1\}}\}$.

The dimension of an VLMC is $Dim_{VLMC}(d) = |\tau|(|A| - 1)$. Typically, $|\tau| \ll |A|^d$.

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

$$p(a|x_{-d:-1}) = p(a|y_{-d:-1}) \text{ if and only if } x_{-d:-1}, y_{-d:-1} \in \mathcal{C}_i.$$

The dimension of an MMM is $Dim_{MMM}(d) = K(|A| - 1)$. Typically, $K \ll |A|^d$.

Yet, in the high-dimensional setting, the model parameters (e.g. the transition probabilities) can be estimated only if $d_n \leq C \log(n)$ for some constant $C > 0$.

Yet, in the high-dimensional setting, the model parameters (e.g. the transition probabilities) can be estimated only if $d_n \leq C \log(n)$ for some constant $C > 0$.

To see this, recall that the estimator of $p(a|x_{-d:-1})$ computed from $X_{1:n}$ is defined as

$$\hat{p}_n(a|x_{-d:-1}) = \frac{N_n(x_{-d:-1}, a)}{\sum_{b \in A} N_n(x_{-d:-1}, b)} = \frac{N_n(x_{-d:-1}, a)}{\bar{N}_n(x_{-d:-1})},$$

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

Yet, in the high-dimensional setting, the model parameters (e.g. the transition probabilities) can be estimated only if $d_n \leq C \log(n)$ for some constant $C > 0$.

To see this, recall that the estimator of $p(a|x_{-d:-1})$ computed from $X_{1:n}$ is defined as

$$\hat{p}_n(a|x_{-d:-1}) = \frac{N_n(x_{-d:-1}, a)}{\sum_{b \in A} N_n(x_{-d:-1}, b)} = \frac{N_n(x_{-d:-1}, a)}{\bar{N}_n(x_{-d:-1})},$$

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

For the estimator $\hat{p}_n(a|x_{-d:-1})$ to have any meaning, we need that $\bar{N}_n(x_{-d:-1}) \geq 1$.

Yet, in the high-dimensional setting, the model parameters (e.g. the transition probabilities) can be estimated only if $d_n \leq C \log(n)$ for some constant $C > 0$.

To see this, recall that the estimator of $p(a|x_{-d:-1})$ computed from $X_{1:n}$ is defined as

$$\hat{p}_n(a|x_{-d:-1}) = \frac{N_n(x_{-d:-1}, a)}{\sum_{b \in A} N_n(x_{-d:-1}, b)} = \frac{N_n(x_{-d:-1}, a)}{\bar{N}_n(x_{-d:-1})},$$

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

For the estimator $\hat{p}_n(a|x_{-d:-1})$ to have any meaning, we need that $\bar{N}_n(x_{-d:-1}) \geq 1$.

By ergodicity, $\bar{N}_n(x_{-d:-1}) \approx n\mathbb{P}(X_{1:d} = x_{-d:-1})$.

Yet, in the high-dimensional setting, the model parameters (e.g. the transition probabilities) can be estimated only if $d_n \leq C \log(n)$ for some constant $C > 0$.

To see this, recall that the estimator of $p(a|x_{-d:-1})$ computed from $X_{1:n}$ is defined as

$$\hat{p}_n(a|x_{-d:-1}) = \frac{N_n(x_{-d:-1}, a)}{\sum_{b \in A} N_n(x_{-d:-1}, b)} = \frac{N_n(x_{-d:-1}, a)}{\bar{N}_n(x_{-d:-1})},$$

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

For the estimator $\hat{p}_n(a|x_{-d:-1})$ to have any meaning, we need that $\bar{N}_n(x_{-d:-1}) \geq 1$.

By ergodicity, $\bar{N}_n(x_{-d:-1}) \approx n\mathbb{P}(X_{1:d} = x_{-d:-1})$. If the transition probabilities are bounded below from zero, then $\exists\ c > 0$ such that $\mathbb{P}(X_{1:d} = x_{-d:-1}) < e^{-cd}$.

Yet, in the high-dimensional setting, the model parameters (e.g. the transition probabilities) can be estimated only if $d_n \leq C \log(n)$ for some constant $C > 0$.

To see this, recall that the estimator of $p(a|x_{-d:-1})$ computed from $X_{1:n}$ is defined as

$$\hat{p}_n(a|x_{-d:-1}) = \frac{N_n(x_{-d:-1}, a)}{\sum_{b \in A} N_n(x_{-d:-1}, b)} = \frac{N_n(x_{-d:-1}, a)}{\bar{N}_n(x_{-d:-1})},$$

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

For the estimator $\hat{p}_n(a|x_{-d:-1})$ to have any meaning, we need that $\bar{N}_n(x_{-d:-1}) \geq 1$.

By ergodicity, $\bar{N}_n(x_{-d:-1}) \approx n\mathbb{P}(X_{1:d} = x_{-d:-1})$. If the transition probabilities are bounded below from zero, then $\exists$ $c > 0$ such that $\mathbb{P}(X_{1:d} = x_{-d:-1}) < e^{-cd}$.

In this case, it follows that we need $1 \leq ne^{-cd}$ implying that $d \leq C \log n$ with $C = 1/c$.

What if $d_n = \beta n$ for some $\beta \in (0, 1)$?

What if $d_n = \beta n$ for some $\beta \in (0, 1)$?

Why $d_n = \beta n$ is important? Many natural phenomena have very long memory!

What if $d_n = \beta n$ for some $\beta \in (0, 1)$?

Why $d_n = \beta n$ is important? Many natural phenomena have very long memory!

In this talk: 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

$$p(a|x_{-d:-1}) = \lambda_0 p_0(a) + \sum_{j=-d}^{-1} \lambda_j p_j(a|x_j),$$

where:

- $p_0(\cdot), p_j(\cdot, b), j \in \{-d, \ldots, -1\}$, $b \in A$ are probability measures on A.
- $\lambda_0, \lambda_1, \ldots, \lambda_{-d} \in [0, 1]$ such that $\sum_{j=-d}^{0} \lambda_j = 1$.

MTD models

Markov chains of order d such that

$$p(a|x_{-d:-1}) = \lambda_0 p_0(a) + \sum_{j=-d}^{-1} \lambda_j p_j(a|x_j),$$

where:

- $p_0(\cdot), p_j(\cdot, b), j \in \{-d, \dots, -1\}$, $b \in A$ are probability measures on A.
- $\lambda_0, \lambda_1, \dots, \lambda_{-d} \in [0, 1]$ such that $\sum_{j=-d}^{0} \lambda_j = 1$.

For each lag $j \in \{-d, \dots, -1\}$, let $\delta_j = \lambda_j \max_{b,c \in A} d_{TV}(p_j(\cdot|b), p_j(\cdot|c))$.

MTD models

Markov chains of order d such that

$$p(a|x_{-d:-1}) = \lambda_0 p_0(a) + \sum_{j=-d}^{-1} \lambda_j p_j(a|x_j),$$

where:

- $p_0(\cdot), p_j(\cdot, b), j \in \{-d, \dots, -1\}$, $b \in A$ are probability measures on A.
- $\lambda_0, \lambda_1, \dots, \lambda_{-d} \in [0, 1]$ such that $\sum_{j=-d}^{0} \lambda_j = 1$.

For each lag $j \in \{-d, \dots, -1\}$, let $\delta_j = \lambda_j \max_{b,c \in A} d_{TV}(p_j(\cdot|b), p_j(\cdot|c))$.

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

MTD models

Markov chains of order d such that

$$p(a|x_{-d:-1}) = \lambda_0 p_0(a) + \sum_{j=-d}^{-1} \lambda_j p_j(a|x_j),$$

where:

- $p_0(\cdot), p_j(\cdot, b), j \in \{-d, \dots, -1\}$, $b \in A$ are probability measures on A.
- $\lambda_0, \lambda_1, \dots, \lambda_{-d} \in [0, 1]$ such that $\sum_{j=-d}^{0} \lambda_j = 1$.

For each lag $j \in \{-d, \dots, -1\}$, let $\delta_j = \lambda_j \max_{b,c \in A} d_{TV}(p_j(\cdot|b), p_j(\cdot|c))$.

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

Note that $p(a|x_{-d:-1}) = p(a|x_\Lambda)$ and $Dim_{MTD}(d) = |\Lambda||A|(|A| - 1) + (|\Lambda| - 1)$.

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 $\beta \in (0, 1)$.

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 $m < n$, $S \subseteq \{-d, \dots, -1\}$, $x_S \in A^S$ and $a \in A$, let

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

In the definition of $\hat{p}_{m,n}(a|x_S)$ 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

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

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

Choice of the random threshold

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

$$s_{m,n}(x_S) = \sqrt{\frac{\alpha(1+\varepsilon)}{2\bar{N}_{m,n}(x_S)}} \sum_{a \in A} \sqrt{V_{m,n}(a, x_S)} + \frac{\alpha|A|}{6\bar{N}_{m,n}(x_S)},$$

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

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

Choice of the random threshold

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

$$s_{m,n}(x_S) = \sqrt{\frac{\alpha(1+\varepsilon)}{2\bar{N}_{m,n}(x_S)}} \sum_{a \in A} \sqrt{V_{m,n}(a, x_S)} + \frac{\alpha |A|}{6\bar{N}_{m,n}(x_S)},$$

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

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

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 $\Lambda \subseteq S$.

From now on, we focus on the case $A = \{0, 1\}$.

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 $S \subseteq \{-d, \ldots, -1\}$ 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 $\Lambda \subseteq S$.

From now on, we focus on the case $A = \{0, 1\}$.

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

Proposition 1. Under Assumption 1 there exists $\kappa > 0$ such that the following property holds: for all $S \subseteq \{-d, \ldots, -1\}$ 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 $X_{1:m}$.

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 $\hat{S}_m$.

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

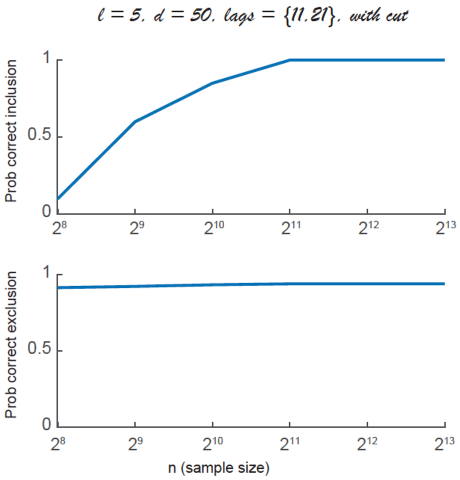
To build $\hat{S}_m$, we do as follows. Fix $0 \leq \ell \leq d$.

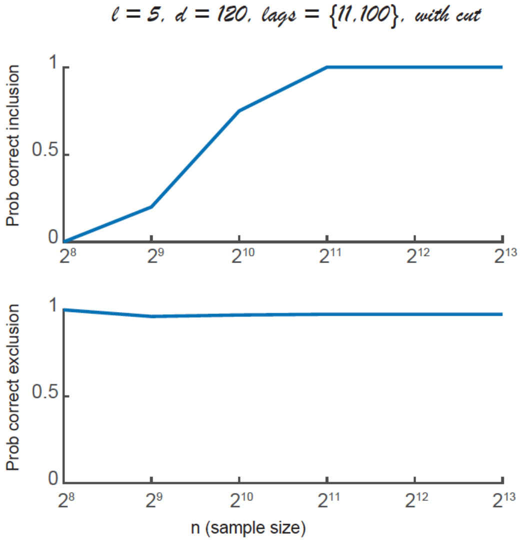
1. Set $\hat{S}_m = \emptyset$.
2. While $|\hat{S}_m| < \ell$, compute $j \in \arg\max_{k \in \hat{S}_m^c} \hat{\nu}_{m,k,\hat{S}_m}$ and include j in $\hat{S}_m$.

Theorem 1. (Consistency) Take $m = n/2$ and assume $d = \beta m$ for some $\beta \in (0,1)$. Suppose Assumption 1 holds and let $\kappa > 0$ given by Proposition 1. Let $\hat{\Lambda}_n$ be the FSC estimator computed with $\ell = 2\kappa^{-2}$ and $\alpha = (1+\eta)\log(n)$ for some $\eta > 0$. Under some other mild assumptions and if $\ell \leq (1-\gamma)/2\log_2(n)$ for some $\gamma \in (0,1)$, then there exits a constant $C > 0$ such that $\mathbb{P}(\hat{\Lambda}_n \neq \Lambda) \to 0$ as $n \to \infty$, as long as

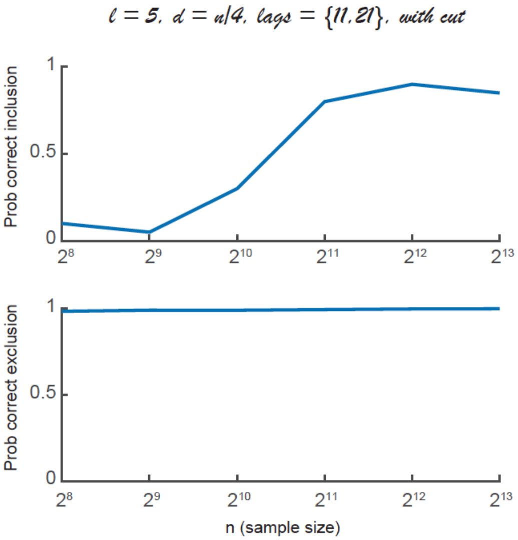
$$\min_{j \in \Lambda} \delta_j^2 \geq C \frac{\log(n)}{n^{(1+\gamma)/2}}.$$

Simulations: FSC estimator





Simulations: FSC estimator



Simulations: transition probability estimation

MTD model used: $p(a|x_{-d:-1}) = \lambda_0 p_0(a|x_0) + \lambda_i p_i(a|x_i) + \lambda_j p_j(a|x_j)$ where $\lambda_0 = 0.2$, $\lambda_i = \lambda_j = 0.4$, $p_i(0|0) = p_i(1|1) = p_j(0|0) = p_j(1|1) = 0.7$.

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

Further theoretical guarantees of FSC estimator

Theorem 2. Take $m = n/2$ and assume $d = \beta m$ for $\beta \in (0,1)$. Suppose $|\Lambda| \leq L$ with L known and the MTD model satisfies some weak dependence conditions. Let $\hat{\Lambda}_n$ be the FSC estimator constructed with parameters $\ell = L$ and $\alpha = (1+\eta)\log(n)$ for $\eta > 0$. Under some other mild assumptions, there exits a positive constant $C > 0$ such that

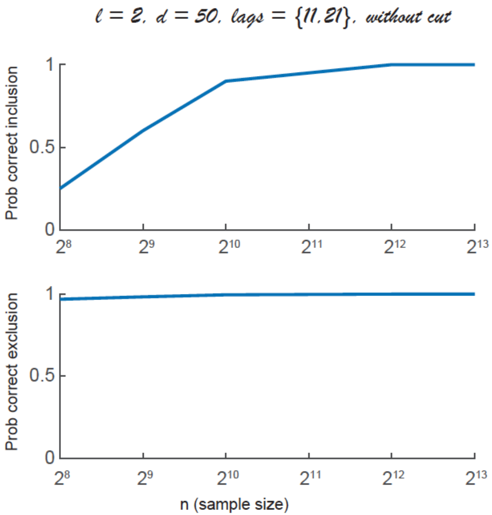
$$\mathbb{P}(\hat{\Lambda}_n \neq \Lambda) \to 0 \text{ as } n \to \infty,$$

as long as

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

If $|\Lambda| = L$ and the MTD satisfies the weak dependence conditions, then we estimate Λ by $\hat{S}_m$. In this case, we neither need the CUT step not to split the data into two pieces!

Simulations: FSC without CUT



Final comments

If in Theorem 1 we suppose also that the Inward weak condition holds, then

$$\kappa = \frac{\Gamma_1 p_{min}^2 \min_{j\in\Lambda} \delta_j}{2\sqrt{|\Lambda|}}.$$

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

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

What about multivariate MTD models?