Markov substitute models and statistical inference in linguistics

Olivier Catoni, joint work with Thomas Mainguy

CNRS, INRIA – CLASSIC
Département de Mathématiques et Applications,
ENS, 45 rue d'Ulm, 75 230 Paris Cedex 05,
`Olivier.Catoni@ens.fr`

Séminaire Parisien de Statistique,
Institut Henri Poincaré,
April 7, 2014

Toric grammars in action

A training sample:

```
1 [0 He is a clever guy .
1 [0 He is doing some shopping .
1 [0 He is laughing .
1 [0 He is not interested in sports .
1 [0 He is walking .
1 [0 He likes to walk in the streets .
1 [0 I am driving a car .
1 [0 I am riding a horse too .
1 [0 I am running .
1 [0 Paul is crossing the street .
1 [0 Paul is driving a car .
1 [0 Paul is riding a horse .
1 [0 Paul is walking .
1 [0 Peter is walking .
1 [0 While I was walking , I saw Paul crossing the street .
```

The estimated toric grammar:

```
10 [0 He likes to walk ]6 ]3 streets .
2 [0 ]1 ]8 clever guy .
2 [0 ]1 doing some shopping .
2 [0 ]1 laughing .
2 [0 ]1 not interested ]6 sports .
2 [0 ]1 riding ]8 horse .
2 [0 ]1 riding ]8 horse ]2 .
2 [0 ]1 running .
24 [0 ]7 am ]5 .
28 [0 Paul is ]5 .
40 [0 He is ]5 .
4 [0 ]1 crossing ]3 street .
4 [0 ]1 driving ]8 car .
5 [0 ]4 is ]5 .
6 [0 ]1 walking .
7 [0 Peter is ]5 .
8 [0 While ]7 was ]5 , ]7 saw ]4 ]5 .
10 [1 He is
2 [1 Peter is
```

```
2 [1 While ]7 was ]5 , ]7 saw ]4
6 [1 ]7 am
8 [1 Paul is
2 [2 too
30 [3 the
14 [4 Paul
1 [4 Peter
16 [5 crossing ]3 street
16 [5 driving ]8 car
16 [5 riding ]8 horse
34 [5 walking
8 [5 ]5 too
8 [5 ]8 clever guy
8 [5 doing some shopping
8 [5 laughing
8 [5 not interested ]6 sports
8 [5 running
20 [6 in
50 [7 I
50 [8 a
```

New sentences discovered:

```
1 [0 Paul is driving a car too .
1 [0 Paul is doing some shopping .
1 [0 Paul is laughing .
1 [0 Paul is riding a horse too .
1 [0 Paul is running too .
1 [0 Paul is running .
1 [0 Paul is not interested in sports too .
1 [0 Paul is not interested in sports .
1 [0 Paul is a clever guy too .
1 [0 Paul is a clever guy .
1 [0 Paul is walking too .
1 [0 Peter is driving a car too .
1 [0 Peter is driving a car .
1 [0 Peter is doing some shopping .
1 [0 Peter is laughing .
1 [0 Peter is riding a horse too .
1 [0 Peter is riding a horse .
1 [0 Peter is running too .
1 [0 Peter is running .
1 [0 Peter is not interested in sports .
```

```
1 [0 Peter is a clever guy .
1 [0 Peter is crossing the street .
1 [0 He is driving a car too .
1 [0 He is driving a car .
1 [0 He is riding a horse too .
1 [0 He is riding a horse .
1 [0 He is running too .
1 [0 He is running .
1 [0 He is not interested in sports too .
1 [0 He is crossing the street too .
1 [0 He is crossing the street .
1 [0 He is walking too .
1 [0 I am driving a car too .
1 [0 I am doing some shopping .
1 [0 I am laughing too .
1 [0 I am laughing .
1 [0 I am riding a horse .
1 [0 I am not interested in sports .
1 [0 I am a clever guy .
1 [0 I am crossing the street too .
1 [0 I am crossing the street .
1 [0 I am walking too .
1 [0 I am walking .
```

```
1 [0 While I was driving a car , I saw Paul doing some shopping too .
1 [0 While I was driving a car , I saw Paul doing some shopping .
1 [0 While I was driving a car , I saw Paul riding a horse .
1 [0 While I was driving a car , I saw Paul crossing the street .
1 [0 While I was driving a car , I saw Paul walking .
1 [0 While I was driving a car , I saw Peter riding a horse .
1 [0 While I was doing some shopping , I saw Paul riding a horse .
1 [0 While I was doing some shopping , I saw Paul walking .
1 [0 While I was laughing too , I saw Peter crossing the street .
1 [0 While I was laughing , I saw Peter riding a horse .
1 [0 While I was riding a horse , I saw Paul driving a car too .
1 [0 While I was riding a horse , I saw Paul driving a car .
1 [0 While I was riding a horse , I saw Paul laughing .
```

```
1 [0 While I was riding a horse , I saw Paul running .
1 [0 While I was riding a horse , I saw Paul walking .
1 [0 While I was riding a horse , I saw Peter not interested in sports .
1 [0 While I was running , I saw Paul laughing .
1 [0 While I was running , I saw Paul not interested in sports .
1 [0 While I was running , I saw Paul a clever guy .
1 [0 While I was running , I saw Paul walking .
1 [0 While I was not interested in sports , I saw Paul driving a car .
1 [0 While I was not interested in sports , I saw Paul riding a horse .
1 [0 While I was a clever guy , I saw Paul running .
1 [0 While I was a clever guy , I saw Paul crossing the street .
1 [0 While I was a clever guy , I saw Paul walking .
1 [0 While I was crossing the street , I saw Paul riding a horse .
1 [0 While I was crossing the street , I saw Paul running .
1 [0 While I was crossing the street , I saw Paul crossing the street .
1 [0 While I was crossing the street , I saw Paul walking .
1 [0 While I was crossing the street , I saw Peter walking .
1 [0 While I was walking , I saw Paul driving a car .
1 [0 While I was walking , I saw Paul laughing .
1 [0 While I was walking , I saw Paul riding a horse .
1 [0 While I was walking , I saw Paul running .
1 [0 While I was walking , I saw Paul not interested in sports .
1 [0 While I was walking , I saw Paul crossing the street too .
1 [0 While I was walking , I saw Paul walking .
1 [0 While I was walking , I saw Peter not interested in sports .
1 [0 While I was walking , I saw Peter walking .
```

Definition of Markov substitute sets

Let D be a dictionary of words, $D^+ = \bigcup_{j=1}^{\infty} D^j$ the set of finite sequences of words and $D^* = \{\epsilon\} \cup D^+$ the set of possibly empty finite sequences of words.

Let $S \in D^+$ be a random sentence, and $\bigl(S_i, 1 \leq i \leq n\bigr)$ a sample of n independent copies of S.

Given a context $x = (x_1, x_2) \in (D^*)^2$, and an expression $y \in D^+$, we define the insertion operator

$$\alpha(x, y) = x_1 \, y \, x_2 \in D^+,$$

that inserts the expression y in the context x.

Definition

A subset $B \subset D^+$ is a Markov substitute set for S when there is a probability measure $q_B \in \mathscr{M}_+^1(B)$ on B (called the substitute measure) such that for any context $x \in (D^*)^2$ and any $y \in B$,

$$\mathbb{P}\big[S = \alpha(x, y)\big] = \mathbb{P}\big[S \in \alpha(x, B)\big]\, q_B(y),$$

where $\alpha(x, B) = \{\alpha(x, y), y \in B\}$.

In simple words, the conditional distribution of y in context x is independent of the context x.
Equivalently, for any $x, x' \in (D^*)^2$, any $y, y' \in B$,

$$\mathbb{P}_S(x_1\, y\, x_2)\mathbb{P}_S(x'_1\, y'\, x'_2) = \mathbb{P}_S(x_1\, y'\, x_2)\mathbb{P}_S(x'_1\, y\, x'_2).$$

(The model could be broaden further by imposing restrictive conditions on the context x.)

Markov chains are a special case of Markov substitute models

If $S = (Z_1, \dots, Z_L)$, where $\big(Z_t, t \in \mathbb{N}\big)$ is a Markov chain, then for any $x = (w_1, w_2) \in D^2$, $\alpha(x, D)$ is a Markov substitute set.

If $S = (Z_1, \dots, Z_\tau)$, where τ is the first hitting time of $C \subset D$, then for any $x = (x_1, x_2) \in (D^+)^2$, any $B \subset (D \setminus C)^+$, $\alpha(x, B)$ is a Markov substitute set.

Basic properties of Markov substitute sets

Any one point set $\{y\}$, $y \in D^+$, is a Markov substitute set.

A subset of a Markov substitute set is itself a Markov substitute set.

If B and C are Markov sets such that $B \cap C \neq \varnothing$, $B \cup C$ is also a Markov substitute set.

The relation

$$y \sim y' \iff \{y, y'\} \text{ is a Markov substitute pair}$$

is an equivalence relation and $D^+/\sim$ forms a partition of D^+ into maximal Markov substitute sets.

Basic properties of Markov substitute sets

The set B is Markov if and only if there is a connected undirected spanning graph $\mathscr{G} \subset B^2$ such that for any $(y, y') \in \mathscr{G}$, $\{y, y'\}$ is a Markov substitute pair.

If $B_j, 1 \leq j \leq \ell$ are Markov substitute sets (including possibly some one point sets), then

$$\gamma(B_1 \dots B_\ell) = \{ s = y_1 \dots y_\ell : y_j \in B_j, 1 \leq j \leq \ell \}$$

is also a Markov substitute set, and

$q_B(y_1 \dots y_\ell) = C_B \prod_{j=1}^{\ell} q_{B_j}(y_j)$, where C_B is a suitable normalizing constant. (The map $(y_1, y_\ell) \mapsto y_1 \dots y_\ell$ may not be one to one, in which case C_B may be different from one!)

Characterization of Markov substitute sets in terms of random parsing

Let $B \subset D^+$ be some subset.

Let us define the set of splits of any sentence $s \in D^+$ as

$$\mathscr{S}(s, B) = \{(x, y), x \in (D^*)^2, y \in B, \alpha(x, y) = s\}.$$

Let us consider some conditional probability kernel $\bigl(\pi(s; x, y), s \in D^+, x \in (D^*)^2, y \in B \cup \{\epsilon\}\bigr)$ such that

$$\mathscr{S}(s, B) \subset \operatorname{supp}\bigl(\pi(s; \cdot)\bigr) \subset \mathscr{S}(s, B) \cup \bigl\{\bigl((s, \epsilon), \epsilon\bigr)\bigr\}.$$

We can for instance take $\pi(s; x, y) = |\mathscr{S}(s, B)|^{-1}$, $(x, y) \in \mathscr{S}(s, B)$.

Let us define the random B-parse X, Y of the random sentence S on the same probability space by its conditional distribution

$$\begin{aligned} &\mathbb{P}_{X,\,Y \mid S\,=\,s}(x, y) \\ &= \begin{cases} \min\limits_{y' \in B} \pi(\alpha(x, y'), x, y'), & (x, y) \in \mathscr{S}(s, B), \\ 1 - \sum\limits_{(x', y') \in \mathscr{S}(s, B)} \mathbb{P}_{X,\,Y \mid S\,=\,s}(x', y'), & x = (s, \epsilon), y = \epsilon. \end{cases} \end{aligned}$$

Lemma

The set B is a Markov substitute set for S if and only if one of the following equations is true

$$\mathbb{P}_{X,\,Y \mid Y \in B} = \mathbb{P}_{X \mid Y \in B} \otimes \mathbb{P}_{Y \mid Y \in B},$$

$$\mathbb{P}_{X \mid Y = y} = \mathbb{P}_{X \mid Y = y'}, \qquad y, y' \in B,$$

$$\mathbb{P}_{Y \mid X = x,\, Y \in B} = \mathbb{P}_{Y \mid Y \in B}$$

Invariant dynamics

Let $(B_j, 1 \le j \le t)$ be a family of Markov substitute sets.

Let us consider a conditional probability kernel $\bigl(\pi(s; x, y, j) : s \in D^+, x \in (D^*)^2, 1 \le j \le t, y \in B_j \cup \{\epsilon\}, \alpha(x, y) = s\bigr)$.

The kernel $\bigl(k(s, s') : s, s' \in D^+\bigr)$, defined as

$$k(s, s') = \begin{cases} \displaystyle\sum_{\substack{x \in (D^*)^2,\\ (y, y') \in (D^*)^2, j}} \pi(s; x, y, j)\, q_{B_j}(y') \left(\frac{\pi(s', x, y', j)}{\pi(s, x, y, j)} \wedge 1 \right), & s \neq s', \\ 1 - \displaystyle\sum_{s'' \in D^+ \setminus \{s\}} k(s, s''), & s' = s, \end{cases}$$

is reversible with respect to $\mathbb{P}_S$.

Proof

$$\mathbb{P}_S(s)k(s,s') = \sum_{\substack{x\in(D^*)^2,\\(y,y')\in(D^*)^2,j}} \mathbb{P}_S\big(\alpha(x,B_j)\big) \\ \times q_{B_j}(y)q_{B_j}(y')\big[\pi(s;x,y,j)\wedge\pi(s';x,y',j)\big] \\ = \mathbb{P}_S(s')k(s',s).$$

Basic properties of reversible dynamics

If $k(y, y') > 0$, $\{y, y'\}$ is a Markov substitute pair and $q_{\{y,y'\}}(y)k(y, y') = q_{\{y,y'\}}(y')k(y', y)$.

For any Markov substitute set B, (including one point sets), $\operatorname{supp}\left(\sum_{t=0}^{\infty} q_B k^t\right)$ is a Markov substitute set.

The communicating classes $\left\{\operatorname{supp}\left(\sum_{t=0}^{\infty} \delta_s k^t\right), s \in D^+\right\}$ forms a partition of D^+ into Markov substitute sets.

Basic properties of reversible dynamics

For any domain $\mathscr{D} \subset D^+$, the reflected dynamics

$$k_{\mathscr{D}}(s, s') = \begin{cases} k(s, s'), & s, s' \in \mathscr{D}, s \neq s', \\ 0, & s' \notin \mathscr{D} \cup \{s\}, \\ 1 - \sum_{s'' \neq s} k(s, s''), & \text{otherwise.} \end{cases}$$

is reversible with respect to $\mathbb{P}_S$.

Test functions

Let us consider a family Θ of subsets $\theta \subset (D^*)^2$, containing all the one point sets $\{x\}$, $x \in (D^*)^2$.

For any pair B_1, B_2 of Markov substitute sets, such that $B_1 \cap B_2 = \varnothing$. let us put $B = B_1 \cup B_2$ and let us consider some B-parse process (X_B, Y_B) and the random variables

$$\begin{aligned} & F_{B_1,B_2,\theta}(X_B, Y_B, p) \\ & \quad = \mathbb{1}\big(X_B \in \theta\big)\big[\mathbb{1}\big(Y_B \in B_1\big) - p\mathbb{1}\big(Y_B \in B\big)\big], \quad \theta \in \Theta, p \in [0,1]. \end{aligned}$$

The set B is a Markov substitute set if and only if there is $p \in [0,1]$ such that for any $\theta \in \Theta$, $\mathbb{E}\big[F_{B_1,B_2,\theta}(X_B, Y_B, p)\big] = 0$. In this case $q_B(B_1) = p$.

Test functions

To estimate $\mathbb{E}\big[F_{B_1,B_2,\theta}(X_B, Y_B, p)\big]$, we can simulate from the sample $\big(S_i, 1 \le i \le n\big)$ an i.i.d. sample $(S_i, X_{B,i}, Y_{B,i})$ such that $\mathbb{P}_{X_{B,i},\, Y_{B,i} | S_i} = \mathbb{P}_{X_B,\, Y_B | S}$, or we can compute

$$
\begin{aligned}
F_{B_1,\,B_2,\,\theta}(s, p) &\overset{\text{def}}{=} \mathbb{E}\big[F_{B_1,B_2,\theta}(X_B, Y_B, \theta) | S = s\big] \\
&= \sum_{x \in (D^*)^2} \sum_{y \in B} \mathbb{1}(x \in \theta) \min_{y' \in B} \pi\big(\alpha(x, y'), x, y'\big) \\
&\qquad \times \big[\mathbb{1}(y \in B_1) - p\mathbb{1}(y \in B)\big] \mathbb{1}\big(s = \alpha(x, y)\big),
\end{aligned}
$$

and consider the i.i.d. samples $F_{B_1,B_2,\theta}(S_i, p)$.

Alternative test functions

Another choice of test functions is

$$
\begin{aligned}
F_{B_1, B_2, \theta}(S, p) = &\sum_{x \in (D^*)^2} \sum_{(y_1, y_2) \in B_1 \times B_2} \mathbb{1}(x \in \theta) \\
&\times \left[\pi\big(\alpha(x, y_1), x, y_1\big) \wedge \pi\big(\alpha(x, y_2), x, y_2\big)\right] \\
&\times \left[\mathbb{1}\big(S = \alpha(x, y_1)\big) q_{B_2}(y_2)(1-p) - \mathbb{1}\big(S = \alpha(x, y_2)\big) q_{B_1}(y_1) p\right] \\
= &\sum_{x \in (D^*)^2} \sum_{y, y' \in (D^*)^2} \pi(S, x, y) \mathbb{1}(x \in \theta) \\
&\times \left[\mathbb{1}(y \in B_1)(1-p) q_{B_2}(y') - \mathbb{1}(y \in B_2) p q_{B_1}(y')\right] \\
&\times \left(\frac{\pi\big(\alpha(x, y'), x, y'\big)}{\pi\big(\alpha(x, y), x, y\big)} \wedge 1\right).
\end{aligned}
$$

Simulations

Consider $\mathbb{P}_{X', Y'|S} = \pi$,
$\mathbb{P}_{Y''|X', Y'} = \mathbb{1}(Y' \in B_1) q_{B_2} + \mathbb{1}(Y' \in B_2) q_{B_1}$,

$$w(X', Y') = \mathbb{E}\left(\left(\frac{\pi(\alpha(X', Y''), X', Y'')}{\pi(\alpha(x', Y'), X', Y')} \wedge 1 \right) \mid X', Y' \right),$$

$\mathbb{P}_{X, Y|X', Y'} =$
$w(X', Y')\delta_{X', Y'} + (1 - w(X', Y'))\delta_{(\alpha(X', Y'), \epsilon), \epsilon}$.

$$F_{B_1, B_2, \theta}(S, p) = \mathbb{E}\Big[\mathbb{1}(X \in \theta) \big[\mathbb{1}(Y \in B_1) - p \mathbb{1}(Y \in B) \big] \mid S \Big].$$

Statistical tests

Let $\mathscr{B}$ be a set of known Markov substitute sets (to start with, we can take $\mathscr{B} = \{\{y,\}, y \in D\}$.

$$\begin{aligned}
\text{Let } C_1 &= \max_{y \in D^+} \sum_{B \in \mathscr{B}} \mathbb{1}\bigl(y \in B\bigr) < \infty, \\
C_2 &= \max_{x \in (D^*)^2} \sum_{\theta \in \Theta} \mathbb{1}\bigl(x \in \theta\bigr) < \infty, \\
h(B, s) &= \mathbb{1}\bigl(\mathscr{S}(s, B) \neq \varnothing\bigr), \\
g(B, \theta, s) &= \mathbb{1}\bigl(\exists (x, y) \in \mathscr{S}(s, B) : x \in \theta\bigr), \\
\nu(B) &= \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{B' \in \mathscr{B}} h(B', S_i) \right)^{-1} h(B, S_i),
\end{aligned}$$

Statistical tests

$$
\begin{aligned}
\xi\bigl(\theta | B_1, B_2\bigr) &= \frac{1}{n} \sum_{i=1}^{n} \Biggl(\sum_{\theta' \in \Theta} g\bigl(B_1 \cup B_2, \theta', S_i\bigr) \Biggr)^{-1} g\bigl(B_1 \cup B_2, \theta, S_i\bigr), \\
\mu(B_1, B_2, \theta) &= \nu(B_1)\nu(B_2)\mu(\theta | B_1, B_2), \\
\overline{h}(s) &= \sum_{B \in \mathscr{B}} h(B, s) \leq C_1\, \ell(s)\bigl(\ell(s) + 1\bigr)/2, \\
\overline{g}(s) &= \sup_{B_1, B_2 \in \mathscr{B}} \sum_{\theta \in \Theta} g\bigl(B_1 \cup B_2, \theta, s\bigr) \leq C_2\, \ell(s)\bigl(\ell(s) + 1\bigr)/2, \\
\mu(B_1, B_2, \theta) &\geq \underbrace{n^{-3} \bigl(\max_{1 \leq i \leq n} \overline{h}(S_i)\bigr)^{-2} \bigl(\max_{1 \leq i \leq n} \overline{g}(S_i)\bigr)^{-1}}_{\substack{\geq 8 n^{-3} C_1^{-2} C_2^{-1} L^{-3} (L+1)^{-3}, \\ \text{where } L = \max_{1 \leq i \leq n} \ell(S_i)}} \mathbb{1}\bigl(\mu(B_1, B_2, \theta) > 0\bigr).
\end{aligned}
$$

Statistical tests

Proposition

Consider some finite set $\Lambda \subset]0,1[$. *With probability at least* $1-2\varepsilon$, *for any* $\lambda \in \Lambda \cup \{-\Lambda\}$, *any* $p \in \mathscr{P} \subset [0,1]$, *any* $\rho \in \mathscr{M}_+^1\big(\mathscr{B}^2 \times \Theta\big)$,

$$\sum_{i=1}^{n} \frac{(k-1)\lambda}{k} \int_{\theta \in \mathscr{B}^2 \times \Theta} F_\theta(S_i,p)\,\mathrm{d}\rho(\theta) - \frac{\lambda^2}{2(1-|\lambda|)^2} \int F_\theta(S_i,p)^2\,\mathrm{d}\rho(\theta)$$

$$\leq \int \sum_{i=1}^{n} \Big[\log\big(1+\lambda F_\theta(S_i,p)\big) - \frac{\lambda}{k} F_\theta(S_i,p)\Big]\,\mathrm{d}\rho(\theta)$$

$$\leq \frac{(k-1)n\lambda}{k} \int \mathbb{E}\big[F_\theta(S,p)\big]\,\mathrm{d}\rho(\theta) + \mathscr{K}(\rho,\mu) + 3\log(k) + \log\big(|\Lambda||\mathscr{P}|/\varepsilon\big).$$

Statistical tests

$$
\begin{aligned}
\text{Let } p(B_1, B_2, \theta) &= \mathbb{P}\bigl(Y_{B_1 \cup B_2} \in B_1 \,\big|\, X \in \theta,\, Y_{B_1 \cup B_2} \in B_1 \cup B_2 \bigr), \\
\text{so that } \mathbb{E}\bigl(F_{B_1, B_2, \theta}(S, p(B_1, B_2, \theta)) \bigr) &= 0, \\
p_+(B_1, B_2) &= \sup \Bigl\{ p(B_1, B_2, \theta) : \theta \in \Theta, \\
& \qquad \mathbb{P}\bigl(X_{B_1 \cup B_2} \in \theta,\, Y_{B_1 \cup B_2} \in B_1 \cup B_2 \bigr) > 0 \Bigr\}, \\
p_-(B_1, B_2) &= \inf \Bigl\{ p(B_1, B_2, \theta) : \theta \in \Theta, \\
& \qquad \mathbb{P}\bigl(X_{B_1 \cup B_2} \in \theta,\, Y_{B_1 \cup B_2} \in B_1 \cup B_2 \bigr) > 0 \Bigr\}, \\
\psi(z) &= \log(1 + z) - z/k,
\end{aligned}
$$

Statistical tests

We will say that (B_1, B_2) is an η-Markov substitute pair of sets when $B = B_1 \cup B_2$ is a Markov substitute set such that $q_B(B_1) \in [\eta, 1-\eta]$. We will say that (B_1, B_2) is a γ-weak η-Markov substitute pair of sets when

$$\eta \leq p_-(B_1, B_2) \leq p_+(B_1, B_2) \leq 1-\eta, \text{ and } p_+(B_1, B_2) - p_-(B_1, B_2) \leq \gamma.$$

Statistical tests

Proposition

Let Λ *be a finite subset of* $]0,1[$*. With probability at least* $1-2\varepsilon$*, for any pair* $(B_1,B_2)\in\mathscr{B}^2$*,*

$$\begin{aligned} B_-\big(p_+(B_1,B_2)\big) &\overset{\text{def}}{=} \sup_{\rho\in\mathscr{M}_+^1(\Theta),\lambda\in\Lambda} \int \sum_{i=1}^n \psi\Big(\lambda F_{B_1,B_2,\theta}\big(S_i,p_+(B_1,B_2)\big)\Big)\,\mathrm{d}\rho(\theta) \\ &\qquad - \mathscr{K}(\rho,\mu_1) - 3\log(k) - \log\left(\frac{|\Lambda|}{\varepsilon\,\nu_1(B_1)\,\nu_1(B_2)}\right) \leq 0 \\ B_+\big(p_-(B_1,B_2)\big) &\overset{\text{def}}{=} \sup_{\rho\in\mathscr{M}_+^1(\Theta),\lambda\in\Lambda} \int \sum_{i=1}^n \psi\Big(-\lambda F_{B_1,B_2,\theta}(S_i,p_-(B_1,B_2))\Big)\,\mathrm{d}\rho(\theta) \\ &\qquad - \mathscr{K}(\rho,\mu_1) - 3\log(k) - \log\left(\frac{|\Lambda|}{\varepsilon\,\nu_1(B_1)\,\nu_1(B_2)}\right) \leq 0 \end{aligned}$$

Statistical tests

Therefore, if we reject the hypothesis that $B_1 \cup B_2$ is a Markov substitute set when

$$\inf_{p\in[0,1]} \max\{B_-(p), B_+(p)\} > 0,$$

the probability of making a false rejection (after testing all pairs in $\mathscr{B}^2$) is at most 2ε.
In the same way we can reject the hypothesis that $(B_1, B_2) \in \mathscr{B}^2$ is an η-Markov substitute pair of sets when

$$\inf_{p\in[\eta,1-\eta]} \max\{B_-(p), B_+(p)\} > 0,$$

with a probability of rejecting one of the true η-Markov pairs (after testing all pairs in $\mathscr{B}^2$), not greater than 2ε.

Statistical tests

With probability at least $1-2\epsilon$, for any γ-weak η-Markov substitute pair,

$$\inf_{p\in[\eta,1-\eta-\gamma]} \max\{B_-(p+\gamma), B_+(p)\} \leq 0.$$

For this test, the proability of false rejection is not greater than 2ε.

Probability of false acceptance

Lemma

For any $p \in [0,1]$, *any* $\lambda \in]-1,1[$, *any* B_1, $B_2 \in \mathscr{B}$, *any* $\theta \in \Theta$, *let* $r(B_1, B_2, \theta) = \mathbb{E}\Big(\mathbb{1}\big(X_{B_1 \cup B_2} \in \theta, Y_{B_1 \cup B_2} \in B_1 \cup B_2\big)\Big)$.
With probability at least $1 - 2\varepsilon$,

$$\sum_{i=1}^{n} \psi\big(\lambda F_{B_1,B_2,\theta}(S,p)\big) \geq \log(\varepsilon) - nr(\theta)\Bigg[\lambda \frac{k-1}{k}\big(p - p(\theta)\big) + \frac{\lambda^2}{1-|\lambda|}\left(\frac{k-1}{k} + \frac{\varphi(k^{-1})}{2k^2}\right)\Big(p(\theta)\big(1-p(\theta)\big) + \big(p-p(\theta)\big)^2\Big)\Bigg],$$

where $\varphi(z) = 2z^{-2}\big(\exp(z) - 1 - z\big)$.

Probability of false acceptance

Let

$$\delta = \frac{1}{n}\log\Big[k^3 n^3 \big(\max_{1\le n\le n} \overline{h}(S_i)\big)\big(\max_{1\le i\le n} \overline{g}(S_i)\big)|\Lambda|\varepsilon^{-2}\Big],$$

$$\chi = \sup_{x\in[(2n)^{-1/2},(2n)^{1/2}]} \inf_{\lambda\in\Lambda} \cosh\Big[\log\Big(\frac{\lambda x}{1-\lambda}\Big)\Big],$$

$$a = \frac{4\chi^2 k}{k-1}\left(1+\frac{\varphi(k^{-1})}{2k(k-1)}\right) \le 4.47\chi^2 \text{ when } k=10,$$

$$b = \frac{(2+\sqrt{2})k}{k-1} \le 3.8 \text{ when } k=10.$$

Probability of false acceptance

Let us assume that there are $B_1, B_2 \in \mathscr{B}$, $\theta_+, \theta_- \in \Theta$ such that $\overline{p}_+ = p(B_1, B_2, \theta_+)$, $\overline{p}_- = p(B_1, B_2, \theta_-)$, $r_+ = r(B_1, B_2, \theta_+)$, and $r_- = r(B_1, B_2, \theta_-)$ are such that

$$r_- \wedge r_+ \geq \frac{16\, k\, \chi^2 \delta}{k-1},$$

$$\overline{p}_+ - \overline{p}_- \geq \sqrt{\frac{a\, \overline{p}_+ (1 - \overline{p}_+) \delta}{r_+}} \left(1 + \frac{a\delta}{r_+} \right) + \frac{b\, \delta}{r_+} + \sqrt{\frac{a \overline{p}_- (1 - \overline{p}_-) \delta}{r_-}} \left(1 + \frac{a\, \delta}{r_-} \right) + \frac{b\, \delta}{r_-}.$$

Probability of false acceptance

With probability at least $1-2\varepsilon$, $\inf_{p\in[0,1]} \max\{B_-(p), B_+(p)\} > 0$, so that the probability of false acceptance of $B_1 \cup B_2$ as a Markov substitute set is at most equal to 2ϵ in this case.

More precisely, with probability at least $1-2\varepsilon$,

$$B_-\left(\overline{p}_+ - \sqrt{\frac{a\overline{p}_+(1-\overline{p}_+)\delta}{r_+}}\left(1+\frac{a\delta}{r_+}\right) - \frac{b\delta}{r_+}\right) > 0,$$

$$B_+\left(\overline{p}_- + \sqrt{\frac{a\overline{p}_-(1-\overline{p}_-)\delta}{r_-}}\left(1+\frac{a\delta}{r_-}\right) + \frac{b\delta}{r_-}\right) > 0.$$

Probability of false acceptance

If we assume now that

$$\overline{p}_+ - 1 + \eta \geq \sqrt{\frac{a\overline{p}_+(1-\overline{p}_+)\delta}{r_+}}\left(1+\frac{a\delta}{r_+}\right)+\frac{b\delta}{r_+},$$

$$\text{or that } \eta - \overline{p}_- \geq \sqrt{\frac{a\overline{p}_-(1-\overline{p}_-)\delta}{r_-}}\left(1+\frac{a\delta}{r_-}\right)+\frac{b\delta}{r_-},$$

$$\begin{aligned}\text{or that } \overline{p}_+ - \overline{p}_- \geq \gamma &+ \sqrt{\frac{a\overline{p}_+(1-\overline{p}_+)\delta}{r_+}}\left(1+\frac{a\delta}{r_+}\right)+\frac{b\delta}{r_+}\\ &+ \sqrt{\frac{a\overline{p}_-(1-\overline{p}_-)\delta}{r_-}}\left(1+\frac{a\delta}{r_-}\right)+\frac{b\delta}{r_-}.\end{aligned}$$

Probability of false acceptance

the false acceptance probability of the test that $(B_1, B_2) \in \mathscr{B}^2$ is an γ-weak η-Markov substitute pair of sets is not greater than 2ϵ.

Building syntax trees

Starting from the obvious family of Markov substitute sets $\mathscr{A}_0 = \{\{w\}, w \in D\}$, and assuming that $\mathscr{A}_k \subset 2^{D^+}$ is already constructed, consider the family of Markov sets $\mathscr{B} = \{\gamma(e), e \in \mathscr{A}_k^+\}$.

We can use the above tests to find out new Markov substitute sets of the form $\gamma(e_1) \cup \gamma(e_2)$, where $e_1, e_2 \in \mathscr{A}_k^+$, and add them to $\mathscr{A}_k$ to form $\mathscr{A}_{k+1}$.

To compute the tests, we need to define a kernel $\big(\pi(s; x, y), s \in D^+, x \in (D^*)^2, y \in \gamma(e_1) \cup \gamma(e_2)\big)$.

Building syntax trees

To do this, we can use two kernels $\big(t(e,e'), e \in \mathscr{A}_{j-1}^{+}, e' \in \mathscr{A}_{j}^{+}, 1 \leq j \leq k, \gamma(e) \subset \gamma(e')\big)$, and $\Big(\overline{\pi}(s,e;x,y), s \in D^{+}, e \in \mathscr{A}_{k}^{+}, s \in \gamma(e), e = \alpha(\overline{x},\overline{y}), \overline{x} \in \big(\mathscr{A}_{k}^{*}\big)^{2}, \overline{y} \in \{e_1, e_2\}, x \in \gamma(\overline{x}_1) \times \gamma(\overline{x}_2), y \in \gamma(\overline{y})\Big)$.

The k th iterate of t, t^k, builds a random syntax tree, and we can put $\pi(s;x,y) = \big(t^k \overline{\pi}\big)(s;x,y)$.

The incremental construction of $\mathscr{A}_k$ can be described by rewriting rules $B \rightarrow e_1$, $B \rightarrow e_2$, where $B \in \mathscr{A}_j$, and $e_1, e_2 \in \mathscr{A}_{j-1}^{+}$, forming a context free grammar.

Estimating the language distribution

If B is a Markov substitute set such that $B \cap \operatorname{supp}(\mathbb{P}_{S}) \neq \varnothing$, then $B \subset \operatorname{supp}(\mathbb{P}_{S})$ and $\mathbb{P}_{S \,|\, S \in B} = q_B$.

Given a collection of Markov substitute sets B_j, $1 \leq j \leq t$ and the above defined reversible dynamics k, we may define the random Markov substitute sets

$$C_i = \operatorname{supp}\biggl(\delta_{S_i} \sum_{j=0}^{\infty} k^j\biggr),$$

and estimate $\mathbb{P}_{S}$ by

$$\widehat{\mathbb{P}} = \frac{1}{n} \sum_{i=1}^{n} q_{C_i},$$

Estimating the language distribution

and consequently $\mathrm{supp}(\mathbb{P}_S)$ by $\bigcup_{i=1}^{n} C_i$.

To compute $\mathbb{1}\left[s \in \mathrm{supp}(\widehat{\mathbb{P}})\right]$ for a given s and answer the question : is s a sentence of the language ? we need to compute $\mathbb{1}(s \in C_i)$. The syntax tree can help here, since

$$\mathbb{1}(s \in C_i) = \mathbb{1}\left[\mathrm{supp}\left(\delta_s \sum_{j=0}^{\infty} k^j\, t^k\right) = \mathrm{supp}\left(\delta_{S_i} \sum_{j=0}^{\infty} k^j\, t^k\right)\right].$$