A Bayesian non-parametric methodology for inferring grammar complexity

Robin J. Ryder

Centre de Recherche en Mathématiques de la Décision, Université Paris-Dauphine

17 December 2020
AppliBugs

Joint work with Lawrence Murray (Uber), Judith Rousseau (Oxford) and Achille Thin (CMAP)

Campbell's monkeys

Cercopithecus campbelli. Photo credit: ALAMY.

Vocalizations

play pause resume stop

Calls made of 7 building blocks: Boom, Hok, Hok-oo, Krak, Krak-oo, Wok, Wok-oo.
Examples:

- Boom Boom Hok-oo Hok-oo Hok-oo Hok-oo Hok-oo Krak-oo
- Hok Hok Hok-oo Krak-oo Krak-oo Wok-oo Wok-oo Wok-oo Krak-oo
- Hok Wok-oo Hok Krak-oo Krak-oo Krak-ooWok-oo Krak-oo Krak-oo Krak-oo Krak-oo Krak-oo Krak-oo Wok-oo Wok-oo Krak-oo Krak-oo Krak-oo Krak-oo

(Campbell data were collected by Sumir Keenan.)

Vocalizations

Linguist and Philos (2014) 37:439–501
DOI 10.1007/s10988-014-9155-7

RESEARCH ARTICLE

Monkey semantics: two 'dialects' of Campbell's monkey alarm calls

Philippe Schlenker · Emmanuel Chemla · Kate Arnold · Alban Lemasson · Karim Ouattara · Sumir Keenan · Claudia Stephan · Robin Ryder · Klaus Zuberbühler

Published online: 28 November 2014

Abstract We develop a formal semantic analysis of the alarm calls used by Campbell's monkeys in the Tai forest (Ivory Coast) and on Tiwai island (Sierra Leone)—two sites that differ in the main predators that the monkeys are exposed to (eagles on Tiwai vs. eagles and leopards in Tai). Building on data discussed in Ouattara et al. (PLoS ONE 4(11):e7808, 2009a; PNAS 106(51): 22026–22031, 2009b) and Arnold et al. (Population differences in wild Campbell's monkeys alarm

PNAS PNAS PNAS

Campbell's monkeys concatenate vocalizations into context-specific call sequences

Karim Ouattara[a,b,c], Alban Lemasson[a,1], and Klaus Zuberbühler[d,1]

[a]Laboratoire EthoS "Ethologie Animale et Humaine," Unité Mixte de Recherche 6552, Centre National de la Recherche Scientifique, Université de Rennes 1, Station Biologique, 35380 Paimpont, France; [b]Centre Suisse de Recherches Scientifiques, Taï Monkey Project, 01 BP1303, Abidjan 01, Côte d'Ivoire; [c]Laboratoire de Zoologie et de Biologie Animale, Université de Cocody, 10 BP770, Abidjan 10, Côte d'Ivoire; and [d]School of Psychology, University of St. Andrews, KY16 9JP Saint Andrews, Scotland

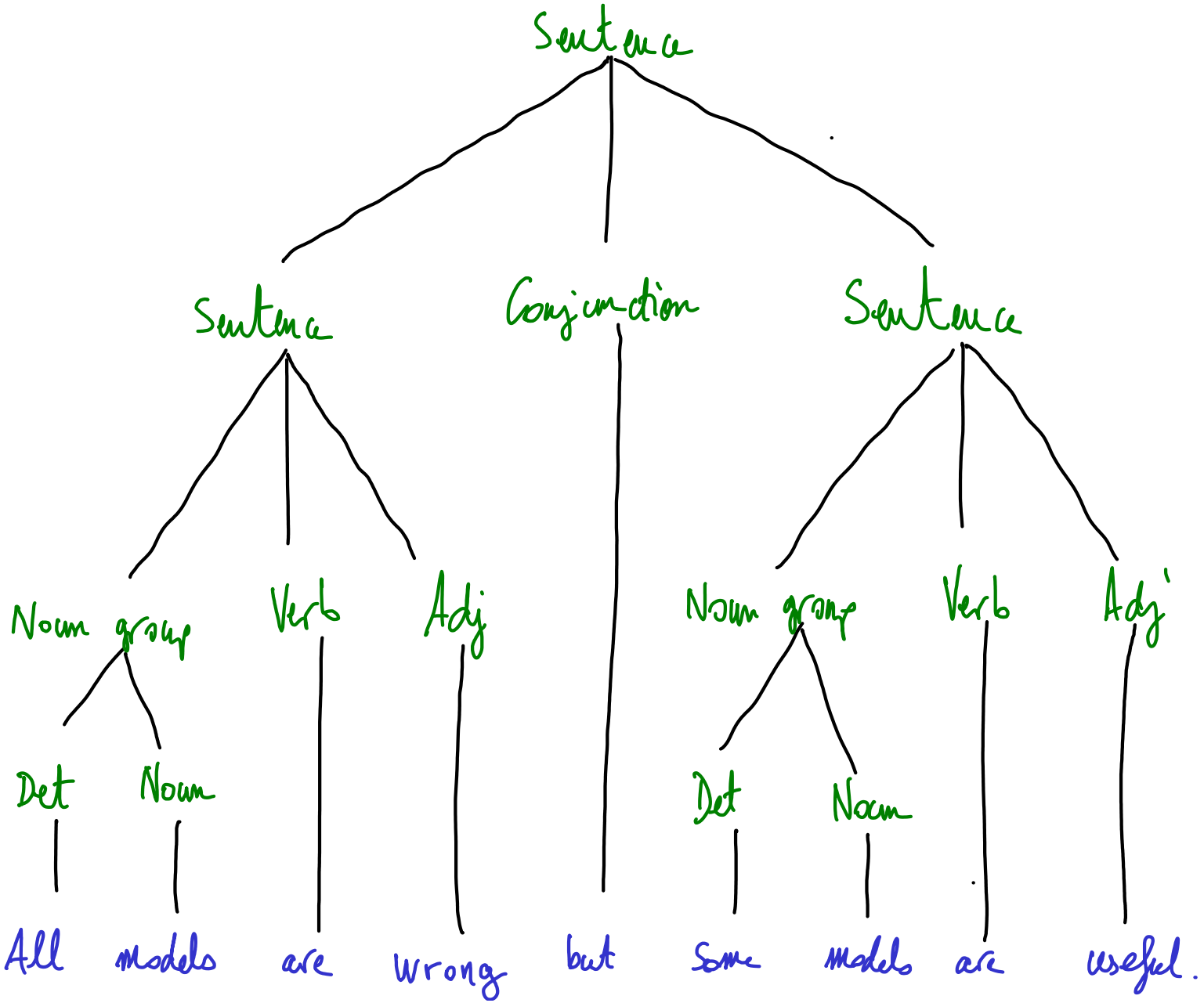
Edited by Charles G. Gross, Princeton University, Princeton, NJ, and approved October 26, 2009 (received for review July 20, 2009)

Primate vocal behavior is often considered irrelevant in modeling human language evolution, mainly because of the caller's limited vocal control and apparent lack of intentional signaling. Here, we present the results of a long-term study on Campbell's monkeys, which has revealed an unrivaled degree of vocal complexity. Adult males produced six different loud call types, which they combined into various sequences in highly context-specific ways. We found stereotyped sequences that were strongly associated with cohesion and travel, falling trees, neighboring groups, nonpredatory animals, unspecific predatory threat, and specific predator classes. Within the responses to predators, we found that crowned eagles triggered four and leopards three different sequences, depending on how the caller learned about their presence. Callers followed a number of principles when concatenating sequences, such as nonrandom transition probabilities of call types, addition of specific calls into an existing sequence to form a different one, or recombination of two sequences to form a third one. We conclude that these primates have overcome some of the constraints of limited vocal control by combinatorial organization. As the different sequences were so tightly linked to specific external events, the Campbell's monkey call system may be the most complex example of 'proto-syntax' in animal communication known to date.

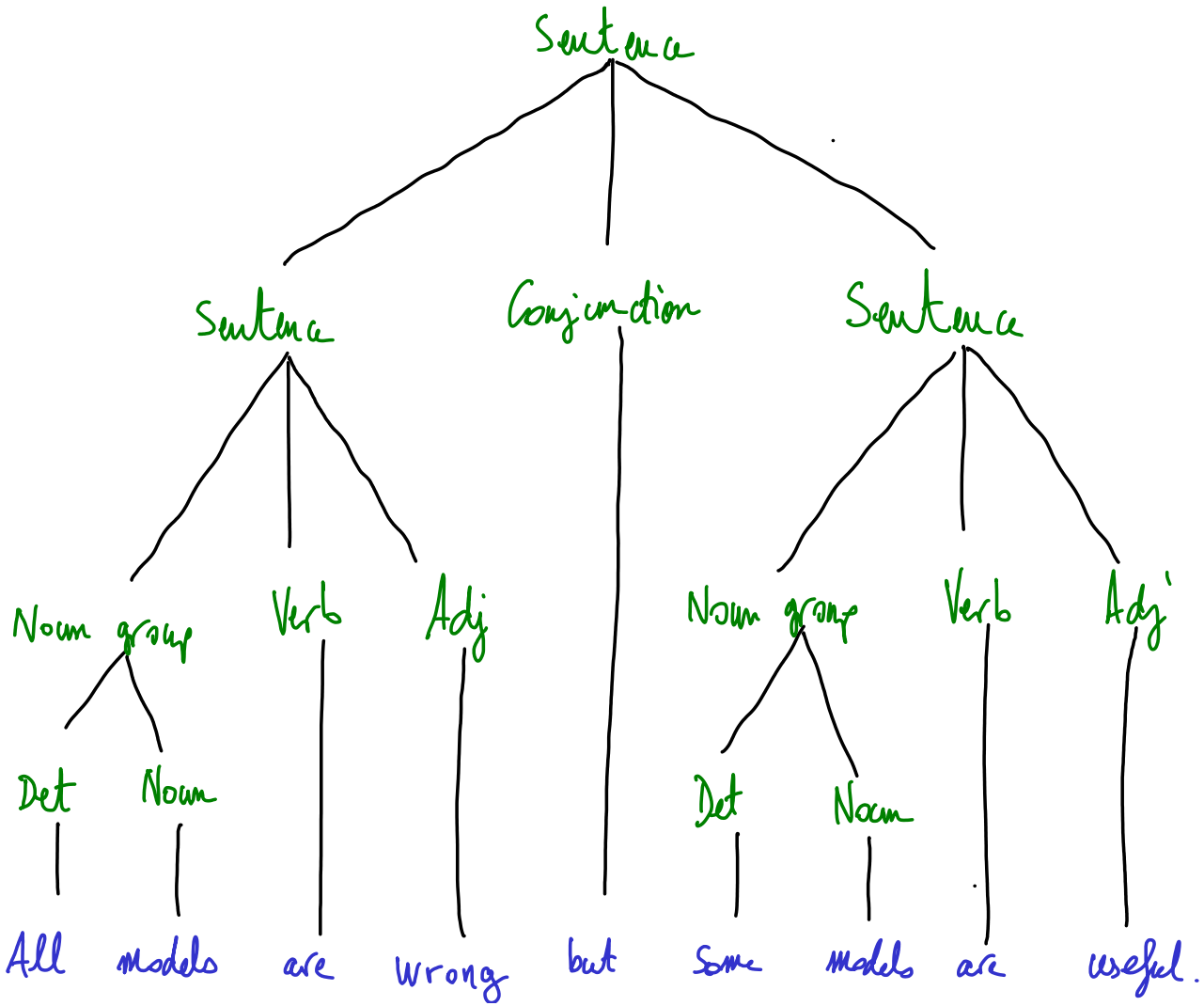
some cases, these structures possess hierarchical organization although very little is known about the relationship betwee acoustic structure and communicative function. A typical findin is that if the structure of a sequence is artificially altered, fo example by changing the composition or order of elements, th signal tends to loose its communicative function (28–30). An other relevant point is that nonhuman primates are perfectl capable of discriminating human speech composed in differen ways [e.g., tamarins (31)] and of comprehending simplifie nonverbal forms of human syntax [e.g., apes (32–35)].

In natural contexts, spontaneous call combinations have als been observed in nonhuman primates, although there are only small number of examples. Chimpanzees combine some of thei calls in nonrandom ways, although the communicative functio of these combinations remains to be investigated (36). Bonobo produce five acoustically distinct call types in response t different foods, with a predictable relationship between th caller's food preference and the relative frequency of the dif ferent calls (37). In putty-nosed monkeys, adult males produc two loud calls, "pyows" and "hacks," in a range of contexts including predation. However, when combining the two calls i one specific way (i.e., a few pyows followed by a few hacks), male

Parse trees


Sentence
Sentence
Conjunction
Sentence
Noun group
Verb
Adj
Noun group
Verb
Adj
Det
Noun
Det
Noun
All
models
are
wrong
but
Some
models
are
useful.

Parse trees



$S \rightarrow S$ Conj S

$S \rightarrow$ NG Verb Adj

NG $\rightarrow$ Det Noun

Det $\rightarrow$ All

Det $\rightarrow$ Some

Noun $\rightarrow$ models

Verb $\rightarrow$ are

Adj $\rightarrow$ wrong

Adj $\rightarrow$ useful

Conj $\rightarrow$ but

A grammar is set of rules

English rules are of the form:

- Sentence $\rightarrow$ NounGroup Verb Adjective
- NounGroup $\rightarrow$ Determiner Noun
- Adjective $\rightarrow$ useful
- ...

Formal grammars

A formal grammar $\mathcal{G}$ is defined by:

- $\mathcal{A} = \{a_1, \ldots, a_K\}$ a set of terminal symbols
- $\mathcal{B} = \{B_1, B_2, \ldots\}$ a set of non-terminal symbols, of which one is specified as the start symbol
- $\mathcal{R}$ a set of rules

Example rule: $r_1 : B_1 \rightarrow a_1 B_2 B_3 a_1$

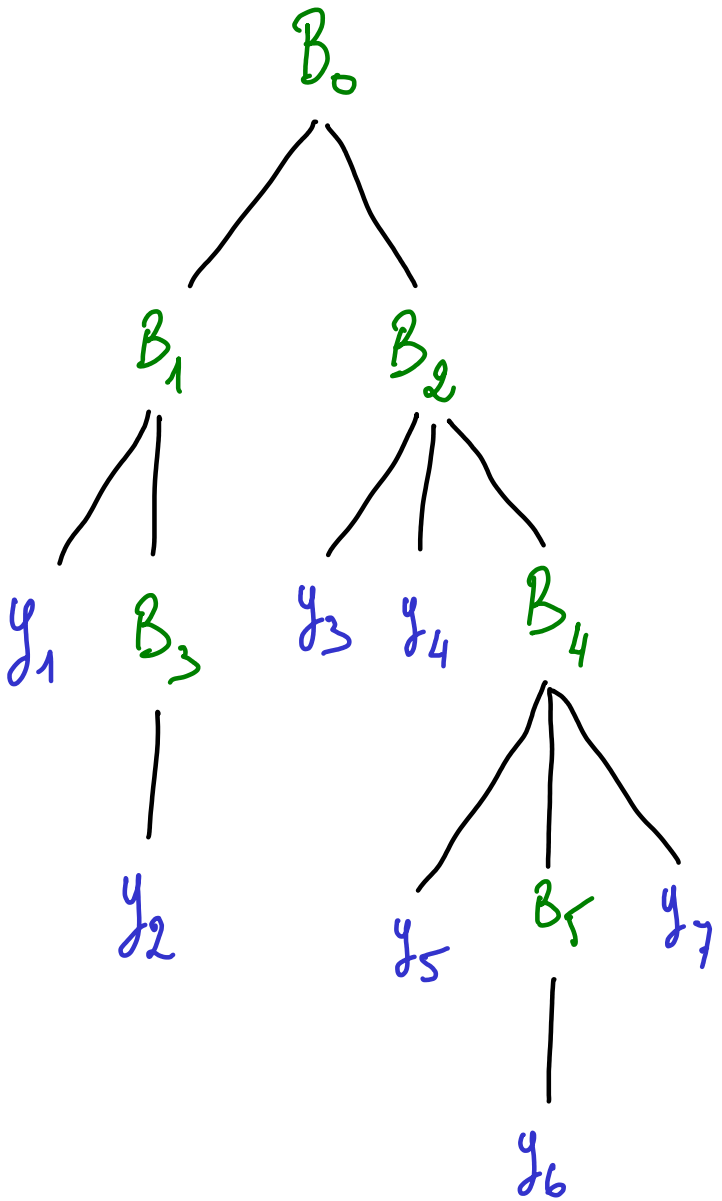
Probabilistic grammars

A sentence is grammatical if it is possible to obtain it by applying a set of rules.

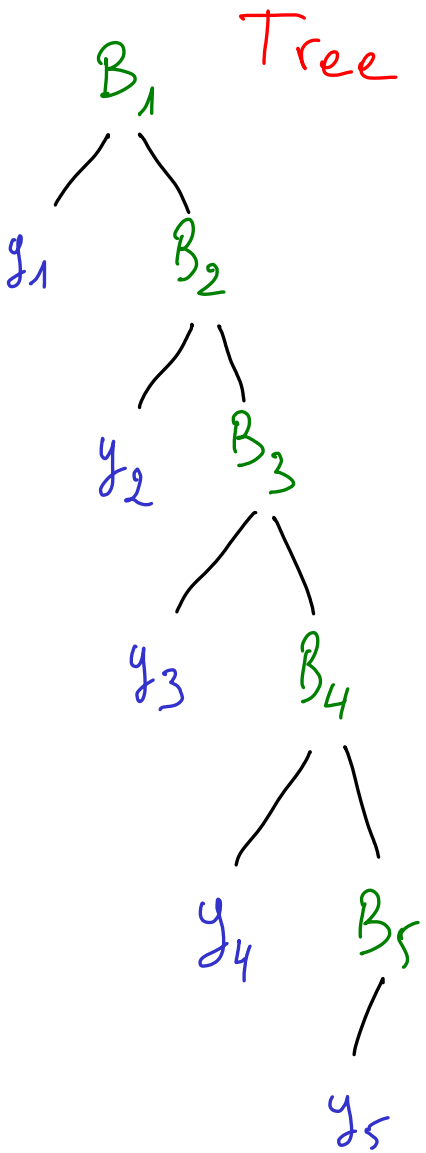
Probabilistic grammars

A sentence is grammatical if it is possible to obtain it by applying a set of rules.
If we assign probabilities to the rules, the grammar becomes *probabilistic*.
With rules+probabilities, we can compute the probability of each sentence.
Since we observe sentences, this defines the likelihood of a set of rules and probabilities.

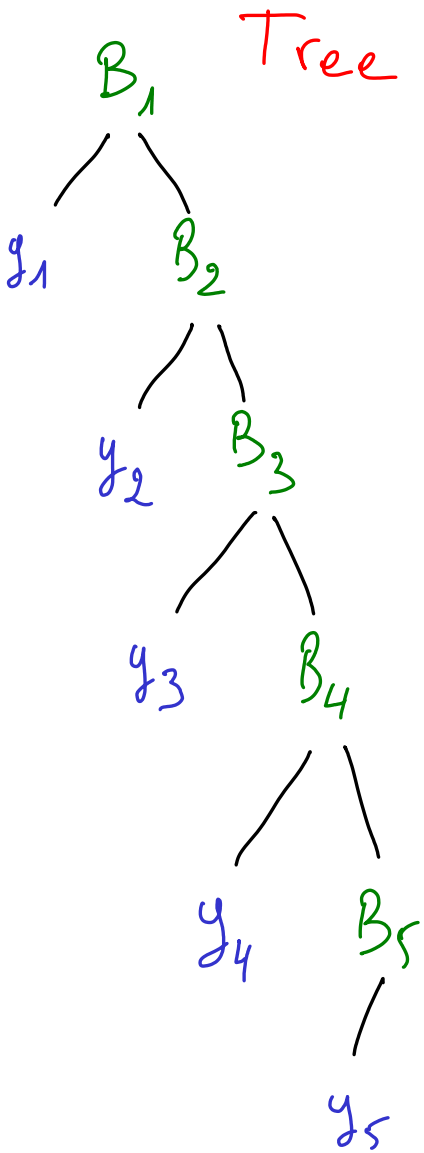
Possible realization of the random tree (1)

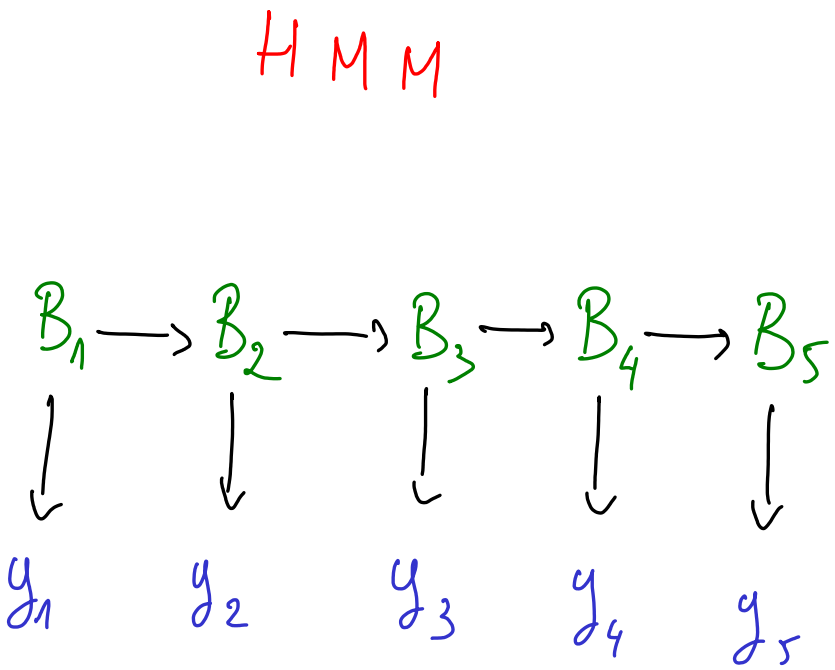


Possible realization of the random tree (2)



Possible realization of the random tree (2)

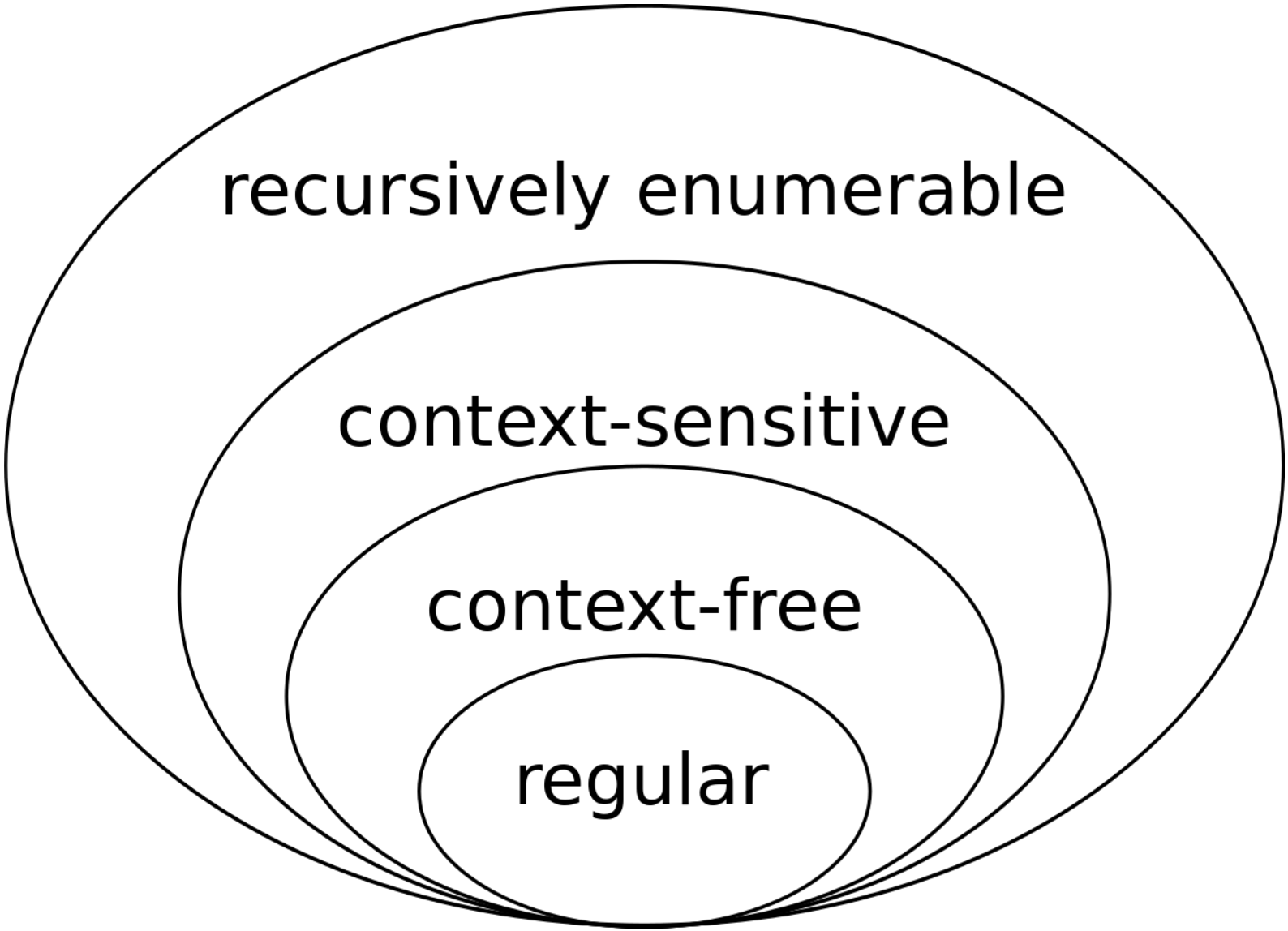




Possible questions

- Given sentences and rules, infer the probabilities
- Given sentences and constraints on the rules, infer the rules (e.g.: constrain to HMMs, infer the hidden states)
- **Given sentences, infer the class of rules**

Chomsky hierarchy



Classes of formal grammars. Figure: J. Filkenstein.

Regular grammar

For a (right) regular grammar, production rules are of the form

- $B_i \rightarrow a_j$
- $B_i \rightarrow a_j B_k$

Regular grammar

For a (right) regular grammar, production rules are of the form

- $B_i \rightarrow a_j$
- $B_i \rightarrow a_j B_k$

This corresponds to a Hidden Markov Model structure.
Key feature: no long-term memory.

Context-free grammar

For a context-free grammar, the production rules are of the form

$$B \to \alpha$$

where α is a string of terminals and non-terminals, for example $B_2 \to B_3 B_4 a_6 a_2 B_3 B_2$.

Context-free grammar

For a context-free grammar, the production rules are of the form

$$B \to \alpha$$

where α is a string of terminals and non-terminals, for example $B_2 \to B_3 B_4 a_6 a_2 B_3 B_2$.
This corresponds to a tree structure.

Context-free grammar

For a context-free grammar, the production rules are of the form

$$B \to \alpha$$

where α is a string of terminals and non-terminals, for example $B_2 \to B_3 B_4 a_6 a_2 B_3 B_2$.
This corresponds to a tree structure.

Any CFG can be rewritten in Greibach Normal Form, where production rules are of the form

- $B_i \to a_j$
- $B_i \to a_j B_k$
- $B_i \to a_j B_k B_\ell$
- $B_i \to a_j B_k B_\ell B_m$
- $B_i \to a_j B_k B_\ell B_m B_n$
- ...

We consider probabilistic versions of these grammars: each production rule is associated with a probability.

Main question

We are given a set of strings output from a formal grammar $\mathcal{G}$. We wish to decide which class of grammars $\mathcal{G}$ belongs to, e.g. test between

$$H_0 : \mathcal{G} \text{ is regular} \quad \textit{vs} \quad H_1 : \mathcal{G} \text{ is context-free.}$$

Main question

We are given a set of strings output from a formal grammar $\mathcal{G}$. We wish to decide which class of grammars $\mathcal{G}$ belongs to, e.g. test between

$$H_0 : \mathcal{G} \text{ is regular} \quad \textit{vs} \quad H_1 : \mathcal{G} \text{ is context-free.}$$

Note that both classes contain an uncountably infinite number of grammars which are compatible with the observations.

Pumping lemma

Let L be the set of sentences legal under $\mathcal{G}$ (i.e. the set of sentences with positive probability). If $\mathcal{G}$ is regular grammar, then (pumping lemma, Rabin & Scott 1959)

$$\exists p \geq 1, \forall s \in L, \text{ if } |s| > p \text{ then } \exists x, y, z \text{ such that } s = xyz$$

with

$$|xy| \leq p \quad |y| \geq 1$$

and all the sentences

$$xyz, xyyz, xyyyz, \ldots, xy^n z, \ldots \in L.$$

How to choose a model: traditional way

- If we can communicate with a speaker of $\mathcal{G}$, we can build an infinite set of sentences, authorized by $\mathcal{G}$, as well as sentences which are impossible under $\mathcal{G}$.
- Use these to build a contradiction to the pumping lemma, thus proving that $\mathcal{G}$ is not regular.
- In English: "an animal ate", "an animal an animal ate ate", "(an animal)^n (ate)^n".

Practical limitations of the pumping lemma

Problem: it is difficult to ask monkeys whether an infinite set of sentences are legal under their grammar.

General case

- No specific information about $\mathcal{G}$.
- Finite number of sentences produced by $\mathcal{G}$.
- Need to explore a number of potential grammars.
- Need to penalize grammars, by number of states and number of rules.

Bayesian setting

- We need to estimate the marginal likelihood for each model
- For this, construction process of CFGs, which defines a prior over grammars, depending on their complexity
- The process we propose corresponds to mutually nested Chinese Restaurant Processes
- Estimation using Sequential Monte-Carlo

Chinese Restaurant Process

A Chinese Restaurant Process is a discrete-time stochastic process; parameter θ.

- Infinite number of tables.
- Customer 1 sits at table 1.
- After n customers, let c_k be the number of customers sitting at table k. Then customer $n+1$ chooses table k wp

$$\frac{c_k}{n+\theta}$$

and a new table wp

$$\frac{\theta}{n+\theta}.$$

"Riches get richer": a table with many customers will be chosen with higher probability by the next customers.

Choosing rules

- For each non-terminal, we have a set of rules.
- These rules come from a Chinese Restaurant Process (one CRP per non-terminal).
- When B_i is at the top of the stack, we generate a rule from the ith CRP.
- The n_ith time that B_i is on the stack, we pick existing rule r_{ij} with probability $\frac{n_{ij}}{n_i+\theta}$.
- We create a new rule with probability $\frac{\theta}{n_i+\theta}$.

Creating a new rule

When we need to create a new rule, it will be of the form $r_{ij} : B_i \to a_k B_{k_1} B_{k_2} \dots B_{k_\ell}$.

- Draw $a_k \sim \mathit{Categorical}(\mathcal{A})$ (Dirichlet prior on probabilities).
- Draw $\ell \sim \mathit{Poisson}(\lambda)$ (in practice, take $\lambda \approx 1$).
- Draw sequentially $B_{k_1}, \dots, B_{k_\ell}$ from a CRP.

We are thus able to create new non-terminals.

Embedding regular grammars

- With this representation, the class of regular grammars is naturally embedded in the class of context-free grammars.
- For a regular grammar, we create new rules with $\ell \in \{0, 1\}$.

Random trees

This corresponds to a latent coloured random tree structure. The model choice question then boils down to: is the latent tree binary (wp 1) or is it of higher arity at at least one node?

Theoretical results (1)

Property 1
The stochastic process (mutually nested CRPs) is well-defined.
(Explicit Hierarchical Dirichlet Process representation.)

For non-terminals:

- $Q \sim DP(M_1, H_0)$, $M_1 > 0$, H_0 a probability distribution on $\mathbb{R}$.
- Sethuraman representation:

$$Q = \sum_{j=1}^{\infty} q_j \delta_{(b_j)}, \quad q_j = V_j \prod_{i<j} (1 - V_i), \quad V_i \overset{iid}{\sim} \text{Beta}(M_1, 1), \quad b_j \overset{iid}{\sim} H_0$$

- Non-terminals: $\mathcal{B} = \{B_0, b_j, j \geq 1\}$.

For non-terminals:

- $Q \sim DP(M_1, H_0)$, $M_1 > 0$, H_0 a probability distribution on $\mathbb{R}$.
- Sethuraman representation:

$$Q = \sum_{j=1}^{\infty} q_j \delta_{(b_j)}, \quad q_j = V_j \prod_{i<j} (1 - V_i), \quad V_i \overset{iid}{\sim} \text{Beta}(M_1, 1), \quad b_j \overset{iid}{\sim} H_0$$

- Non-terminals: $\mathcal{B} = \{B_0, b_j, j \geq 1\}$.

For rules (of the form $B \to aB_1 \cdots B_L$):

- For each $B \in \mathcal{B}$ generate independently $P_B \sim DP(M_2, H_P(\cdot|Q))$, where $H_P(\cdot|Q)$ is a probability distribution on $\mathcal{A} \times \cup_{n=0}^{\infty} \mathcal{B}^n$ defined by $H_P(\beta = (a, B_1 \cdots B_L)|Q) = \mu_{\mathcal{A}}(a) H_L(L) \left(\prod_{j=1}^{L} Q(B_j) \right)^{\mathbb{I}_{L>0}}$
- Generate rules R_b (in the form $b \to \beta$) by $R_b^i | P_b \overset{iid}{\sim} P_b, i \geq 1$

We can integrate out the P_bs and then Q to recover the $\mathbb{Q}$ the CRP associated to the model $B_i|Q \overset{iid}{\sim} Q$, and $Q \sim DP(M_1 H_0)$ so that

$$\mathbb{Q}(dB_n|B_1, \cdots, B_{n-1}) = \frac{M_1}{M_1 + n - 1} H_0(dB_n) + \frac{1}{M_1 + n - 1} \sum_{i=1}^{n-1} \delta_{(B_i)}(dB_n)$$

Theoretical results (2)

Property 2

$$P_{H_1}(H_0) = 0.$$

Theoretical results (2)

Property 2

$$P_{H_1}(H_0) = 0.$$

This property is a necessary (but not sufficient) condition for Bayes factors to converge.

Theoretical results (3)

Property 3
Almost surely, a random regular grammar generates only finite sentences.
Almost surely, a random context-free grammar generates both finite and infinite sentences.

$$\Pi_0 \left(P[\textit{infinite sentence}|\mathcal{G}] = 0 \right) = 1$$

$$\Pi_1 \left(0 < P[\textit{infinite sentence}|\mathcal{G}] < 1 \right) = 1$$

Implementation

The model is implemented in the probabilistic programming language Birch (`www.birch-lang.org`, Murray et al. 2017).
We draw from the joint posterior using Sequential Monte-Carlo (SMC), which allows us to estimate the marginal likelihood of each model.

Results

We run this procedure on:

- Synthetic regular grammars
- Synthetic context-free grammars
- Alarm calls from Campbell monkeys
- Ongoing: parts-of-speech tags from English literature
- In the future: music sheets

Model choice

- Since the models are embedded, a natural idea for model comparison would be to simulate from the posterior in the complex context-free model, and see whether the credible intervals include the simpler regular model.
- In practice, this does not work: we take data from a known regular grammar, and simulate from the posterior in the context-free model. The rules generated are not regular (we get $I > 1$).
- Bayes factors are thus the best option.

Bayesian model choice

Standard method for model choice in a Bayesian setting: compute the Bayes factor

$$BF = \frac{m_0(y)}{m_1(y)}$$

where

$$m_i(y) = \int L(\theta; y)\pi(\theta)\, d\theta$$

is the normalizing constant of the posterior.

Bayesian model choice

Standard method for model choice in a Bayesian setting: compute the Bayes factor

$$BF = \frac{m_0(y)}{m_1(y)}$$

where

$$m_i(y) = \int L(\theta; y)\pi(\theta)\, d\theta$$

is the normalizing constant of the posterior.
Interpretation: if $BF > 1$, we favour model 0; if $BF < 1$, we favour model 1. Jeffreys' scale of evidence:

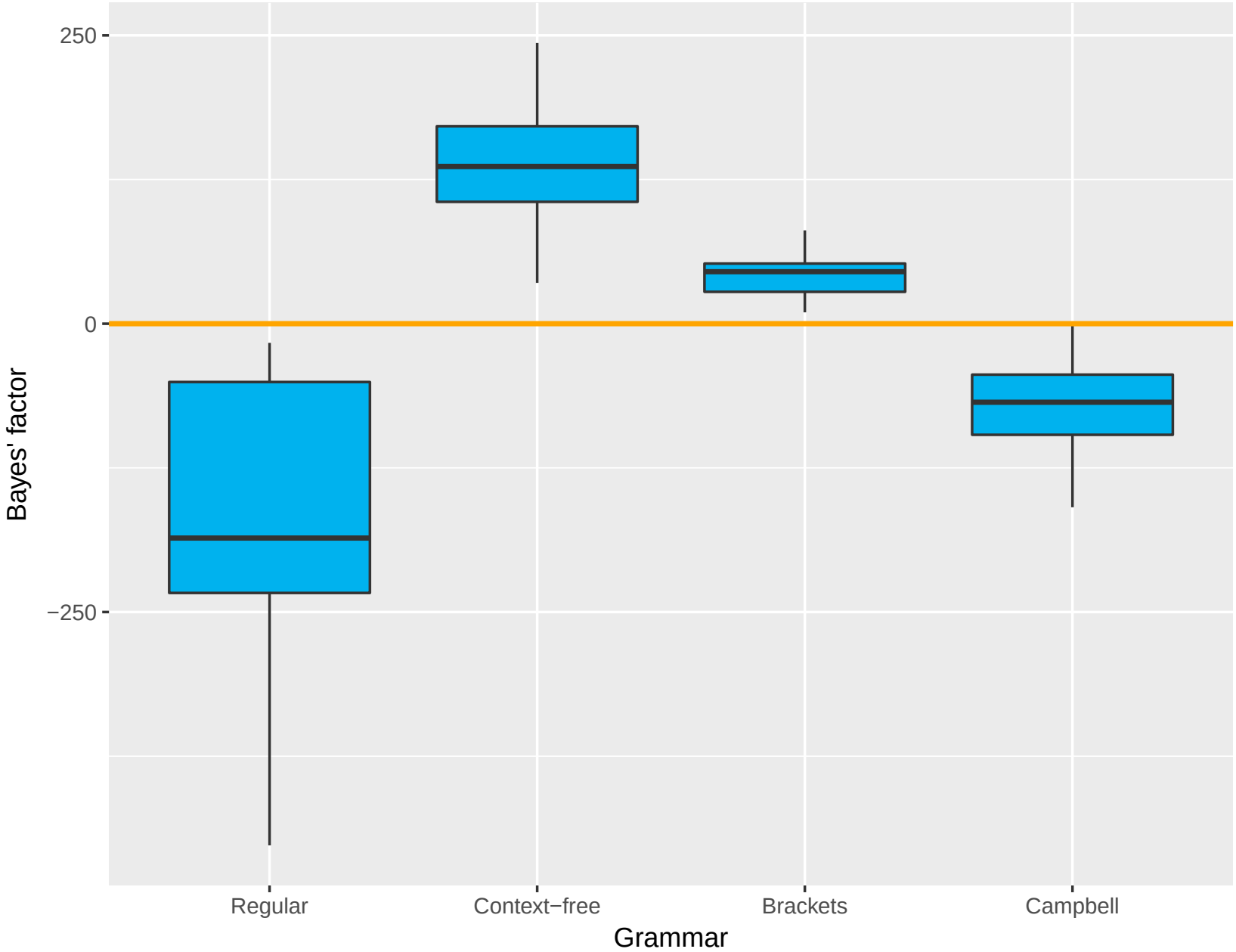
- if $0 < \log_{10}(BF) < \frac{1}{2}$: weak evidence in favour of model 0
- if $\frac{1}{2} < \log_{10}(BF) < 1$: substantial evidence in favour of model 0
- if $1 < \log_{10}(BF) < 2$: strong evidence in favour of model 0
- if $2 < \log_{10}(BF)$: decisive evidence in favour of model 0

and symmetrically in favour of model 1 if $\log_{10}(BF) < 0$.

Number of particles necessary to decide the sign of the Bayes factor

- An SMC with N particles gives an estimate $\hat{M}_i^{(N)}$ of $M_i = \log m_i(y)$ ($i = 0, 1$). Estimator is consistent, with variance $\propto \frac{1}{N}$.
- We are interested in deciding the sign of $M_0 - M_1$.
- Ad hoc procedure: obtain 5 realizations $\hat{M}_0^{(N)}$ and of $\hat{M}_1^{(N)}$. If all 25 elements of $\left(\hat{M}_{0,j}^{(N)} - \hat{M}_{1,k}^{(N)}\right)_{j,k}$ are of the same sign, use that to choose the model, else increase N.
- In other words, if $\min_k \hat{M}_{1,k}^{(N)} > \max_j \hat{M}_{0,j}^{(N)}$ choose model 1; if $\min_j \hat{M}_{0,j}^{(N)} > \max_k \hat{M}_{1,k}^{(N)}$ choose model 0; else increase N.
- Typical number of particles necessary is $10^3 - 10^5$.

Results


250
0
−250
Bayes' factor
Regular
Context−free
Brackets
Campbell
Grammar

Future work

- Influence of the choice of Greibach Normal Form: we are currently checking that the procedure is valid even when the grammar can be written in a sparse form in Chomsky Normal Form, but not sparsely in the Greibach Normal Form.
- For computational reasons, we are limited to small numbers of terminal symbols and medium-sized data sets.
- Extending to choosing between other classes in the Chomsky hierarchy.

Questions


© Txema Garcia Laseca / Comedy Wildlife Photography Awards 2019