Context tree selection and linguistic rhythm retrieval from written texts

Antonio Galves, Charlotte Galves,
Nancy L. Garcia, Florencia Leonardi

July 8, 2009

Abstract

We introduce a new criterion to select in a consistent way the probabilistic context tree generating a sample. The basic idea is to construct a totally ordered set of candidate trees. This set is composed by the "champion trees", the ones that maximize the likelihood of the sample for each number of degrees of freedom. The smallest maximizer criterion selects the infimum of the subset of champion trees whose gain in likelihood is negligible. This study was motivated by the linguistic challenge of retrieving rhythmic patterns from written texts. Applied to a data set consisting of texts extracted from daily newspapers, our algorithm identifies different context trees for European Portuguese and Brazilian Portuguese. This is compatible with the long standing conjecture that European Portuguese and Brazilian Portuguese belong to different rhythmic classes. Moreover, these context trees have several interesting properties which are linguistically meaningful.

1 Introduction

In this paper we address the challenging linguistic question of how to retrieve rhythmic patterns from written texts. The search for rhythmic signatures in written texts is important from different scientific point of views. For example, it is an important ingredient for developing realistic text to speech synthesizers. Also, it is a helpful tool to describe the historical evolution of the rhythm of a natural language, as the only available evidence is that which can be retrieved from written texts.

Stochastic chains with memory of variable length appear as good candidates to model the symbolic chains obtained by encoding written texts in natural languages. In effect, it can be argued on linguistic grounds that in a rhythmic chain each new symbol is a probabilistic function of a suffix (ending string) of the string of past symbols. Moreover, the length of the relevant portion of the past depends on the past itself. This corresponds precisely to the class of *probabilistic context tree models* introduced by Rissanen in his seminal 1983 paper in which the relevant part of the past is called a *context*.

Given a finite realization of a stochastic chain with memory of variable length, the basic statistical question is how to identify the smallest probabilistic context tree fitting the data. This issue has been addressed by an increasing number of papers, starting with Rissanen (1983) who introduced the so-called *algorithm Context* to perform this task. See also Ron et al. (1996), Bühlmann and Wyner (1999), Galves et al. (2008) and Galves and Löcherbach (2008) for a survey. We refer also the reader to Busch et al. (2009) for a recent application of context trees to classification protein families.

A different approach was proposed by Csiszár and Talata (2006) who showed that context trees can be consistently estimated in linear time using the *Bayesian Information Criteria* (BIC). We refer the reader to this paper for a nice description of other approaches and results in this field, including the *Context Tree Weighting Method* (CTW) introduced by Willems et al. (1995). See also Garivier (2006) for a general discussion of the field.

Both the algorithm Context and the BIC procedure requires the specification of some constants. For the algorithm Context, the constant appears in the threshold used in the pruning decision. For the BIC, the constant appears in the penalization term. In both cases, the consistency of the algorithm does not depend on the specific choice of the constant. However, for finite samples - even with very large size - the choice of the constant does matter. Different constants will give different answers ranging from the maximum tree (constant close to zero) to the root tree (constant very large).

An adaptive procedure to choose the asymptotic context tree from a finite sample is a most important question from the point of view of applied statistics. This is achieved by the smallest maximizer criterion introduced in the present paper. In informal terms, the criterion selects the tree which is the infimum of a subset of the set of "champion trees". We rigorously prove that the smallest maximizer criterion selects in a consistent way the finite context tree generating the infinite sample.

Now the question is how to apply this criterion to identify the tree from a finite sample. We propose a new algorithm based on resampling to implement the criterion. We make a simulation study which indicates the suitability of the procedure.

We apply the smallest maximizer criterion and its implementation to solve a long standing linguistic problem. Can we retrieve rhythmic features from written texts?

Modern Portuguese provides an interesting case to be analyzed from the point of view of rhythm. European Portuguese and Brazilian Portuguese (henceforth EP and BP respectively) share the same lexicon. From the point of view of external language, they also produce a great number of superficially identical sentences (for the dichotomy *internal and external language* we refer the interested reader to Chomsky 1985). However EP and BP have been argued to implement different *rhythms* (cf. for instance Révah 1958, Sândalo et al. 2006 and Frota and Vigário 2001 for a critical discussion of the rhythmic features of BP and EP). We refer the reader to Ramus (2002) for an illuminating discussion of the rhythmic class conjecture and to Cuesta et al. (2007) for a statistical classification of speech data according to their rhythmic features.

In the present paper, the smallest maximizer criterion was applied to a real linguistic

data set, constituted for the needs of the present study, with randomly chosen written texts extracted from a corpus of Brazilian and European Portuguese daily newspapers. These texts were encoded using a finite set of labels, expressing a few basic rhythmic features which can be retrieved automatically from written texts.

The smallest maximizer criterion selects different context trees for BP and EP. The difference between the context trees can be linguistically interpreted in a way which is compatible with current hypotheses on the characteristic features of the different rhythmic classes.

This article is organized as follows. Section 2 presents the class of probabilistic context tree models and states the main theoretical results supporting the proposed algorithm. Section 3 presents the smallest maximizer criterion (SMC) and its implementation is given in Section 4. Section 5 is dedicated to the linguistic case study which is the original motivation for this article. In Section 6 a simulation study illustrates in a concrete way the good performance of the algorithm implementing the smallest maximizer criteria. A final discussion is presented in Section 7. The mathematical proof of the theorems are given in Appendix 1. Appendix 2 discusses the preprocessing of the linguistic data and the computation of the degrees of freedom of the models.

2 Stochastic chains with memory of variable length

Let A be a finite alphabet. We will use the shorthand notation w_m^n to denote the string $(w_m, \ldots, w_n)$ of symbols in the alphabet A. The length of this string will be denoted by $\ell(w_m^n) = n - m + 1$. We say that a sequence s_{-j}^{-1} is a *suffix* of a sequence w_{-k}^{-1} if $j \leq k$ and $s_{-i} = w_{-i}$ for all $i = 1, \ldots, j$. This will be denoted as $s_{-j}^{-1} \preceq w_{-k}^{-1}$. If $j < k$ then we say that s is a proper suffix of w and denote this relation by $s \prec w$. The same definition applies when $w_{-\infty}^{-1}$ is a semi-infinite sequence.

Definition 2.1. *A finite subset τ of $\cup_{k=1}^{\infty} A^{\{-k,\ldots,-1\}}$ is an* irreducible tree *if it satisfies the following conditions.*

1. Suffix property. *For no $w_{-k}^{-1} \in \tau$ we have $w_{-k+j}^{-1} \in \tau$ for $j = 1, \ldots, k-1$.*
2. Irreducibility. *No string belonging to τ can be replaced by a proper suffix without violating the suffix property.*

It is easy to see that the set τ can be identified with the set of leaves of a rooted tree with a finite set of labeled branches. Elements of τ will be denoted either as w or as w_{-k}^{-1} if we want to stress the number of elements of the string.

Let $p = \{p(\cdot|w) \colon w \in \tau\}$ be a family of probability measures on A indexed by the elements of τ. The elements of τ will be called *contexts* and the pair (τ, p) will be called *probabilistic context tree*. The number of contexts in τ will be denoted by $|\tau|$. The height $\ell(\tau)$ of the tree τ is the maximal length of a context in τ, that is

$$\ell(\tau) = \max\{\ell(w) \colon w \in \tau\}.$$

We recall that we are assuming that τ is a finite set and therefore ℓ is finite.

Definition 2.2. *The stationary ergodic stochastic process (X_t) on A has memory of variable length compatible with the probabilistic context tree (τ, p) if*

1. *For any $n \geq \ell(\tau)$ and any sequence x_{-n}^{-1}*

$$\mathbb{P}(X_0 = a \mid X_{-n}^{-1} = x_{-n}^{-1}) = p(a|x_{-j}^{-1}), \quad \text{for all } a \in A, \tag{2.3}$$

   where x_{-j}^{-1} is the only suffix of x_{-n}^{-1} belonging to τ.

2. *No proper suffix of x_{-j}^{-1} satisfies (2.3).*

Definition 2.4. *Define the following partial ordering on the set of all context trees. We will say that $\tau \preceq \tau'$, if for every $v \in \tau'$, there exists $w \in \tau$ such that $w \preceq v$. As usual, whenever $\tau \preceq \tau'$ with $\tau \neq \tau'$ we will write $\tau \prec \tau'$.*

3 Smallest maximizer criterion

Given a finite sample $X_1, \ldots, X_n$ of elements in A generated by (τ^*, p^*), the model selection problem is to find a procedure based on X_1^n to estimate the tree τ^*.

For any finite string w_{-j}^0 with $j \leq d(n)$, we denote by $N_n(w_{-j}^0)$ the number of occurrences of the string w_{-j}^0 in the sample

$$N_n(w_{-j}^0) = \sum_{t=d(n)+1}^{n} \mathbf{1}\{X_{t-j}^t = w_{-j}^0\}\ , \tag{3.1}$$

where $d(n)$ is a suitable function of n such that $d(n) \to \infty$ as $n \to \infty$.

Assuming the sample was generated by a stationary chain, for any finite string w_{-k}^{-1} such that $\sum_{b \in A} N_n(w_{-k}^{-1} b) > 0$, the maximum likelihood estimator of the transition probability $\mathbb{P}(X_0 = a | X_{-k}^{-1} = w_{-k}^{-1})$ is given by

$$\hat{p}_n(a|w_{-k}^{-1}) = \frac{N_n(w_{-k}^{-1} a)}{\sum_{b \in \mathcal{A}} N_n(w_{-k}^{-1} b)}\,, \tag{3.2}$$

where $w_{-k}^{-1} a$ denotes the string $(w_{-k}, \ldots, , w_{-1}, a)$, obtained by concatenating w_{-k}^{-1} and the symbol a.

The likelihood function for a tree τ is given by

$$L_\tau(X_1^n) = \prod_{w \in \tau} \prod_{a \in \mathcal{A}} \hat{p}_n(a|w)^{N_n(wa)}. \tag{3.3}$$

Let $\mathcal{T}_n = \mathcal{T}(X_1, \ldots, X_n)$ be the set of all irreducible trees τ such that

- $\ell(\tau) \leq d(n)$;
- for all $w \in \tau$, $\sum_{b \in A} N_n(wb) > 0$;
- any sequence u with $\sum_{b \in A} N_n(ub) > 0$ has a suffix that belongs to τ or is a suffix (proper or not) of an element in τ.

Let $\mathrm{df}\colon \mathcal{T}_n \to \mathbb{N}$ be a function that assigns to each tree $\tau \in \mathcal{T}_n$ the number of degrees of freedom of the model corresponding to the context tree τ. The definition of $\mathrm{df}(\tau)$ depends on the class of models considered. Without any restriction $\mathrm{df}(\tau) = (|A| - 1)|\tau|$. However, in many scientific data sets we know beforehand that some transitions are not allowed by the nature of the problem. That is the case of the linguistic data set we are considering in our case study presented in Section 5. In general, we can define an incidence function $\chi : \cup_{j=1}^{\infty} A^{\{-j,\ldots,-1,0\}} \to \{0,1\}$ which indicates in a consistent way which are the possible transitions. By consistent we mean that if $\chi(w_{-j}^{-1}a) = 0$ for some w_{-j}^{-1} and $a \in A$ then $\chi(w_{-k}^{-1}a) = 0$ for all $k \geq j$. In this case,

$$\mathrm{df}(\tau;\chi) = \sum_{w\in\tau}\sum_{a\in A}\chi(wa).$$

Obviously, we are using the convention that $\chi(wa) = 0$ means that the transition from w to a is not allowed.

Then

$$\mathcal{T}_n = \bigcup_{g\in\mathcal{G}_n} \mathcal{T}_n^g,$$

where $\mathcal{G}_n = \mathrm{df}(\mathcal{T}_n)$ and $\mathcal{T}_n^g = \{\tau \in \mathcal{T}_n \colon \mathrm{df}(\tau) = g\}$.

For each $g \in \mathcal{G}_n$ let τ_n^g be the tree belonging to the class $\mathcal{T}_n^g$ which maximizes the likelihood of the sample, that is

$$\tau_n^g = \arg\max_{\tau\in\mathcal{T}_n^g} \log L_\tau(X_1^n).$$

Denote by $\mathcal{C}_n$ the class of champion trees belonging to $\mathcal{T}_n$, that is

$$\mathcal{C}_n = \{\tau_n^g \colon g \in \mathcal{G}_n \text{ such that } L_{\tau_n^{g'}}(X_1^n) < L_{\tau_n^g}(X_1^n) \text{ whenever } g' < g\}.$$

Observe that it is possible to have $g' < g$ with $L_{\tau_n^{g'}}(X_1^n) > L_{\tau_n^g}(X_1^n)$. In the definition of $\mathcal{C}_n$ we discard the bigger tree since the tree with less parameters provides larger likelihood.

Define also the class $\mathcal{C}$ of all champion trees for the infinite sample, that is

$$\mathcal{C} = \bigcup_{n\geq 1} \mathcal{C}_n.$$

Observe that the set of all context trees is not totally ordered with respect to the ordering introduced in Definition 2.4. It turns out that, for any n, the set of champion trees $\mathcal{C}_n$ is totally ordered and contains the tree generating the sample for sufficiently large sample sizes. This is the basis for the selection principle and is the content of the next theorem.

Theorem 3.4. *Assume $X_1, \ldots, X_n$ is a sample of an ergodic stochastic process compatible with (τ^*, p^*), with τ^* finite. Then, $\mathcal{C}_n$ is totally ordered with respect to the order $\prec$ and eventually almost surely $\tau^* \in \mathcal{C}_n$ as $n \to \infty$.*

The next theorem is the basis for the smallest maximizer criterion. It shows that there is a change of regime in the gain of likelihood at τ^*.

Theorem 3.5. *Assume $X_1, \ldots, X_n$ is a sample of an ergodic stochastic process compatible with (τ^*, p^*) with τ^* finite. Then, the following results hold eventually almost surely as $n \to \infty$.*

(1) *For any $\tau \in \mathcal{C}_n$, with $\tau \prec \tau^*$, there exists a constant $c(\tau^*, \tau) > 0$ such that*

$$\log L_{\tau^*}(X_1^n) - \log L_\tau(X_1^n) \geq c(\tau^*, \tau)\, n.$$

(2) *For any $\tau \prec \tau' \in \mathcal{C}_n$, with $\tau^* \preceq \tau$, there exists a constant $c(\tau, \tau') \geq 0$ such that*

$$\log L_{\tau'}(X_1^n) - \log L_\tau(X_1^n) \leq c(\tau, \tau') \log n.$$

Theorems 3.4 and 3.5 lead to the following *Smallest Maximizer Criterion.*

Smallest Maximizer Criterion. Select the smallest tree $\hat{\tau}$ in the set of champion trees $\mathcal{C}$ such that

$$\lim_{n\to\infty} \frac{\log L_\tau(X_1^n) - \log L_{\hat{\tau}}(X_1^n)}{n} = 0$$

for any $\tau \succeq \hat{\tau}$.

The next theorem states the consistency of this criterion.

Theorem 3.6. *Let $X_1, X_2, \ldots$ be an ergodic chain compatible with the probabilistic context tree (τ^*, p^*) with τ^* finite. Then,*

$$\mathbb{P}(\hat{\tau} \neq \tau^*) = 0.$$

To avoid technical details and facilitate the reading, we delay the proofs of Theorems 3.4, 3.5 and 3.6 to Appendix 1.

The problem now is how to identify this smallest tree. A procedure doing this is presented in the next section.

4 Implementing the smallest maximizer criterion

In order to select the model first we need an algorithm to compute the set of champion trees $\mathcal{C}_n \subset \mathcal{T}_n$. To do this we explore the relationship between our criteria and the BIC context tree selection.

Definition 4.1. *The BIC context tree estimator with penalizing constant $c > 0$ is defined as*

$$\hat{\tau}_{\text{BIC}}(X_1^n; c) \;=\; \arg\max_{\tau\in\mathcal{T}_n}\{\log L_\tau(X_1^n) - c \cdot df(\tau) \cdot \log n\} \tag{4.2}$$

where $L_\tau(X_1^n)$ is the likelihood of the tree τ given the sample X_1^n and $df(\tau)$ denotes the number of degrees of freedom of the model corresponding to the context tree τ.

Proposition 4.3. *The set of champion trees $\mathcal{C}_n$ is the image of the map*

$$c \in [0, +\infty) \mapsto \hat{\tau}_{\text{BIC}}(X_1^n, c) \in \mathcal{T}_n.$$

Remark: Csiszár and Talata (2006) prove the consistency of the BIC selection procedure in the case of unbounded trees when the length of a candidate context is bounded above by $d(n) = o(\log n)$. Besides the consistency of the procedure, this condition also implies that the estimation can be done in linear time using the context tree maximizing (CTM) algorithm introduced by Willems et al. (1995). Assuming that the tree is bounded, Garivier (2006a) proves consistency of the BIC selection procedure for any diverging function $d(n)$. This is the case we consider here. Therefore, the above proposition implies that all champion trees $\mathcal{C}_n$ can be obtained using the CTM algorithm by changing the penalizing constant in the BIC.

The next step is to identify a tree $\hat{\tau}$ belonging to $\mathcal{C}_n$ for n sufficiently large but finite. Theorem 3.4 guarantees that $\tau^* \in \mathcal{C}_n$. In this case we have to choose, among the champion trees belonging to $\mathcal{C}_n$, the smallest one for which the gain in likelihood is negligible when compared to bigger ones.

To do this, we propose a bootstrap procedure. To determine the change of regime we compare the bootstrap confidence intervals. In practice, we compare the ratio between the gain in log-likelihood and the size of the sample with the boxplots of the resamples. We expect that for $\tau_n^g \succ \tau^*$ the confidence intervals constructed with the increasing sample sizes will decrease, whereas for $\tau_n^g \prec \tau$ the confidence intervals will either converge to a point or increase.

Bootstrap Procedure:

1. For different sample sizes $n_1 < n_2 < \ldots < n_R < n$ obtain B independent bootstrap resamples of $X_1, \ldots, X_n$. Denote these resamples by $\mathbf{X}^{*,(b,j)} = \{X_i^{*,(b,j)}, i = 1, \ldots, n_j\}$ where $b = 1, \ldots, B$ and $j = 1, \ldots, R$.

2. For $j = 1, 2, \ldots, R$ and for all $\tau_n^g \in \mathcal{C}_n$ and its successor $\tau_n^{g'} \in \mathcal{C}_n$ in the $\prec$ order, compute the 1st and 3rd quartile for the ratio

$$\frac{\log L_{\tau_n^g}(\mathbf{X}^*(b,j)) - \log L_{\tau_n^{g'}}(\mathbf{X}^*(b,j))}{n_j}.$$

   Denote them by $Q_1^g(j)$ and $Q_3^g(j)$ respectively.

3. Select the tree $\hat{\tau}$ as the first champion tree τ_n^g such that the resampled confidence interval $[Q_{1,j}^g, Q_{3,j}^g]$ shrinks to zero as j increases.

In Step 1 above, any bootstrap resampling method for stochastic chains with memory of variable length can be used. In our specific case, we use a remarkable feature for our data set, that is, the fact that one of the symbols is a renewal point. This makes it possible to sample randomly with replacement independent strings between two successive renewal points.

5 A linguistic case study

The data we analyze is an encoded corpus of newspaper articles. The electronic files with these articles are available through the project AC/DC (Acesso a Corpora/Disponibilização

de Corpora) at the URL `www.linguateca.pt/acesso`, corpus CHAVE (see Santos and Rocha 2005 for a presentation of the corpus). This corpus contains all the 365 editions of the years 1994 and 1995 from the daily newspapers *Folha de São Paulo* (Brazil) and *O Público* (Portugal). Our sample consists of 80 articles randomly selected from the 1994 and 1995 editions. We chose 20 articles from each year for each newspaper. We ended up with a sample of 97,750 symbols for Brazilian Portuguese (BP) and 105,326 symbols for European Portuguese (EP).

Encoding was made by assigning one of four symbols to each syllable of the text according to whether it is stressed or not and whether it is the beginning of a prosodic word or not. Using the base 2 representation of the integers, this double Boolean classification can represented by the four symbols alphabet $\{0, 1, 2, 3, \}$ where

- 0 = non-stressed, non prosodic word initial syllable;
- 1 = stressed, non prosodic word initial syllable;
- 2 = non-stressed, prosodic word initial syllable;
- 3 = stressed, prosodic word initial syllable.

In the above classification, by *prosodic word* we mean a lexical word together with the functional non stressed words which precede it (cf. for instance Vigário 2003). We recall the fact that in Portuguese the stressed syllable of a word is lexically defined. In other terms, the knowledge of the stressed syllable is part of the knowledge of the word. The orthography of both BP and EP is made in such a way that stressed syllables can be retrieved automatically from the way a stressed word is written.

Additionally we assigned an extra symbol (4) to encode the end of each sentence. Let us call $A = \{0, 1, 2, 3, 4\}$ the alphabet obtained in this way.

This encoding can be performed automatically after a preprocessing of the texts (see Appendix 2). A software written in Perl was developed for this purpose and it is available upon request. The corpora with the encoded newspapers texts can be freely downloaded for academic purposes at URL `www.ime.usp.br/~tycho/smc/data`.

This way to encode written texts according to its rhythmic properties is new, and the data set we are considering has never been analysed from this point of view before.

It is worth observing that the symbolic chain obtained this way is constrained both by general restrictions on possible sentences and by the morphology of Portuguese. Therefore several transitions are impossible. These restrictions are the key to computing the number of degrees of freedom of the proposed model. The full set of restrictions is described in Appendix 2.

To implement the smallest maximizer criterion as described in Section 4 we first need to identify the set of champion trees. By Proposition 4.3 this can be done by changing the penalization constant in the BIC for context trees. As proved in Csiszár and Talata 2006, this can be done in linear time. This way we obtain the set of champion trees for both languages, $\mathcal{C}_{BP}$ and $\mathcal{C}_{EP}$, which ranges from the root tree (independent case), with $|A| - 1$ degrees of freedom, to trees with several thousand degrees of freedom. The function df, which

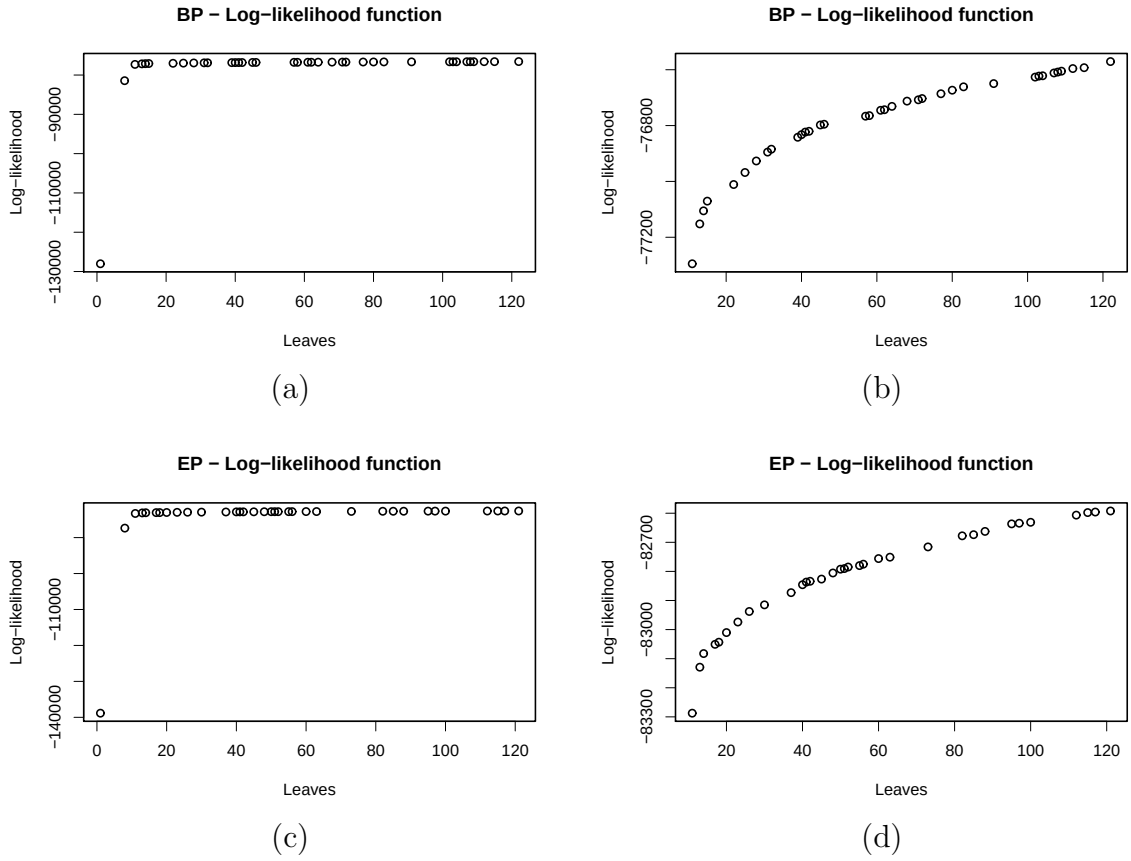


Figure 1: Log-likelihood of the sample as a function of the number of leaves for Brazilian Portuguese (a) and European Portuguese (c). Figures (b) and (d) are zoomed pictures from (a) and (c) for number of leaves bigger than 10.

assigns to each model its number of degrees of freedom, was computed taking into account the constraints of the symbolic chain mentioned in the last paragraph. Figure 1 presents the log-likelihood corresponding to each champion tree for BP and EP according to the number of leaves. The figure clearly suggests that there is a change of regime in a certain region. However, a visual inspection is not enough to detect precisely in which tree it takes place.

To implement the bootstrap procedure we choose three different sample sizes $n_1 = 10,000$, $n_2 = 40,000$ and $n_3 = 70,000$. To each sample size, we resample $B = 250$ times. To resample, we take advantage of a striking feature which is present in all the champion trees. The symbol 4 appears as a renewal point, that is, $p^*(\cdot|w4u) = p^*(\cdot|w4)$ for any finite sequence w. Therefore, we use the independent blocks between two consecutive symbols 4's to perform the usual Efron's independent with replacement bootstrap procedure. The final resample of size n_j is obtained by the concatenation of the successively chosen independent blocks truncated at size n_j.

A software written in C was developed to implement both the identification of the champion trees and the bootstrap procedure. This software is available upon request.

According to the smallest maximizer criterion we compared the ratio of the log-likelihood

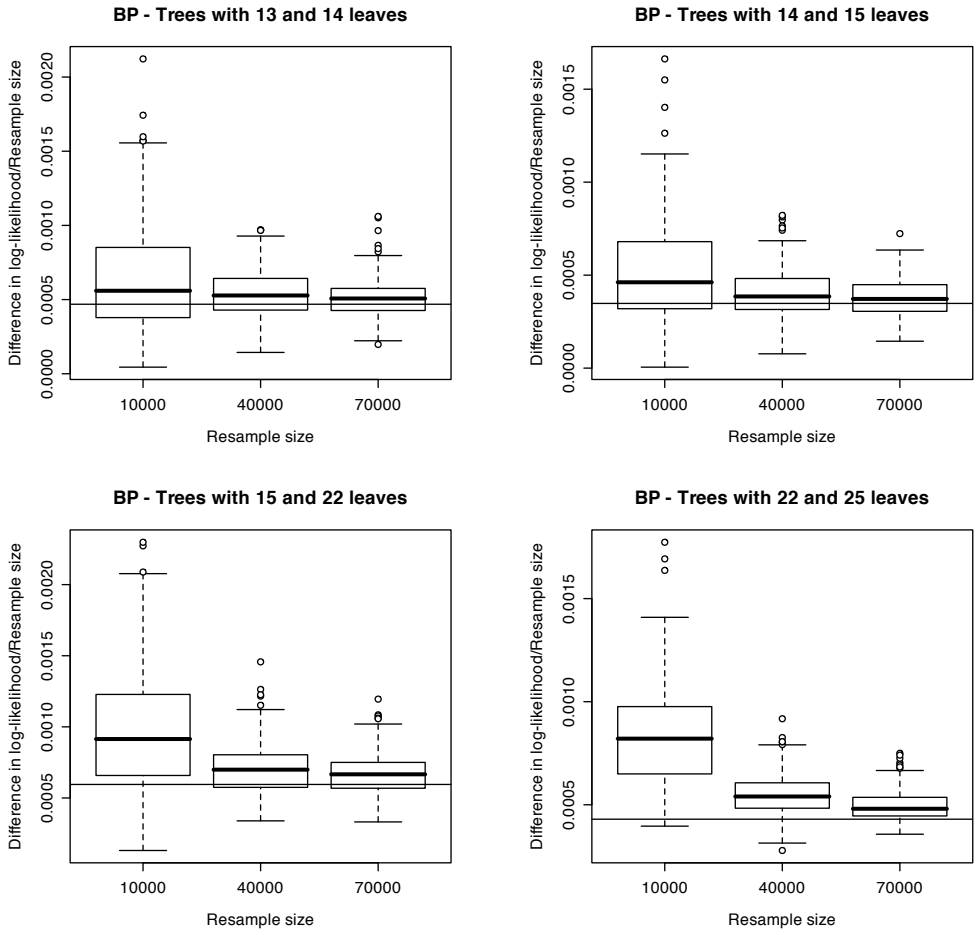


Figure 2: Resampling boxplots of the ratio of the log-likelihood and the sample size for Brazilian Portuguese and observed sample value (solid line).

and the sample size for the resamples. Figures 2 and 3 show some of the corresponding boxplots as well as the observed value for the whole sample (solid line). We can see clearly a change of regime in the tree with 15 leaves for BP and 18 leaves for EP corresponding to trees shown in Figures 4 and 5.

Besides discriminating EP and BP, the selected trees have properties which are linguistically interpretable. First, in both trees, 4 is a context. This is expected on linguistic grounds since both in syntax and in phonology, the sentence is the higher domain.

Second, in both trees, non stressed internal syllables provide poor information about the future. Three successive symbols zero are needed to constitute a context. This is also a welcome result from a linguistic point of view since non stressed non initial syllables do not play a salient role in rhythm by their own, but only by constituting prosodic domains (feet) with stressed syllables, or by being aligned with higher prosodic domains (words or phrases).

Note that a stressed syllable alone is not enough to predict the next symbol. The tables of transition probabilities (Figures 4 and 5) show that in both languages the distribution of what follows a stressed syllable is dependent on the presence or absence of a preceding

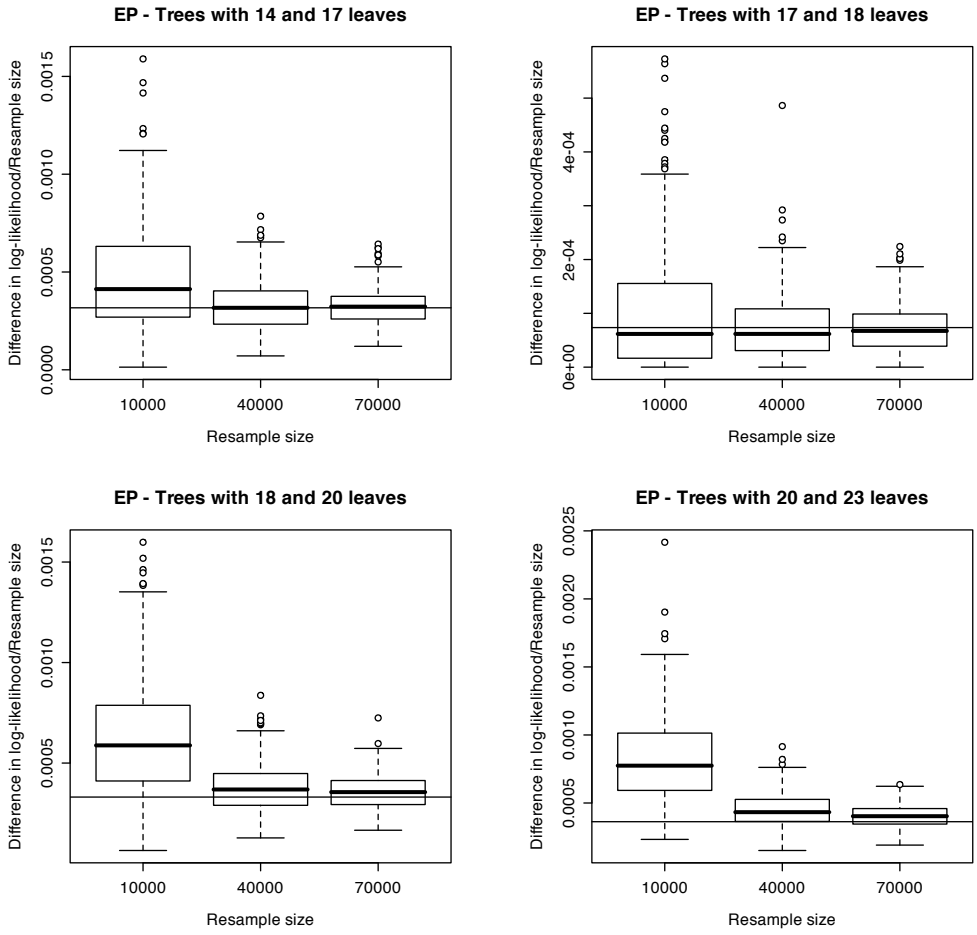


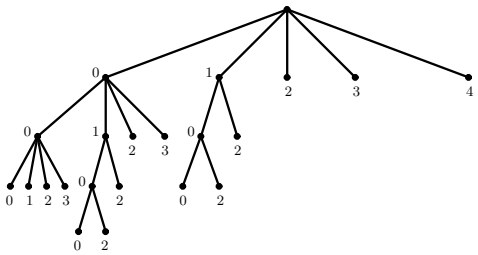
Figure 3: Resampling boxplots of the ratio of the log-likelihood and the sample size for European Portuguese.

prosodic word boundary in the two preceding steps. This fact, arguably derivable from the morphology of Portuguese, does not discriminate EP and BP. This is not surprising since to a great extent EP and BP share the same lexicon.

Finally, according to the selected trees, the main difference between the two languages is that whereas in BP, both 2 (unstressed boundary of a prosodic word) and 3 (stressed boundary of a prosodic word) are contexts, in EP only 3 is. This means that in EP, but not in BP, the words which begin with a stressed syllable behave differently from the words which don't. This is again a welcome result since it is compatible with already observed differences involving the prosodic properties of words in the two languages (cf. Vigário (2003) and Sândalo et al. (2006), among others).

6 Simulation results

We perform a simulation study using the context tree and the transition probabilities presented in Figure 6. We simulate a sample with 100,000 symbols, obtaining 1,882 phrases

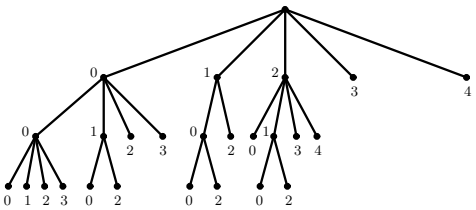


w	$p(0\|w)$	$p(1\|w)$	$p(2\|w)$	$p(3\|w)$	$p(4\|w)$
000	0.29	0.71	0.00	0.00	0.00
100	0.00	0.00	0.67	0.21	0.12
200	0.40	0.60	0.00	0.00	0.00
300	0.00	0.00	0.67	0.22	0.11
0010	0.03	0.00	0.67	0.20	0.10
2010	0.07	0.00	0.66	0.19	0.08
210	0.08	0.00	0.63	0.22	0.07
20	0.45	0.55	0.00	0.00	0.00
30	0.07	0.00	0.64	0.25	0.04
001	0.62	0.00	0.27	0.08	0.03
201	0.72	0.00	0.19	0.07	0.02
21	0.73	0.00	0.18	0.08	0.01
2	0.60	0.40	0.00	0.00	0.00
3	0.69	0.00	0.21	0.10	0.00
4	0.00	0.00	0.66	0.34	0.00

Figure 4: Probabilistic context tree for Brazilian Portuguese.

(sequences delimited by the symbol 4). Using this sample, we estimate the sequence of champion trees by increasing the value of the penalizing constant and considering only trees with height smaller or equal 7. This procedure gives a sequence of trees containing the true tree of 13 leaves. As an illustration of Theorem 3.5, we plot the log-likelihood corresponding to each tree as a function of the number of leaves. We can see a change of regime at the tree with 13 leaves, but its identification using this graphical representation is difficult because the tree with 11 leaves has a log-likelihood very close to the real one.

In order to identify the true tree we asses the convergence of the difference in log-likelihood of two adjacent trees when divided by the sample size. For each size $n_j = j \cdot 10,000$, with $j = 1, 2, \ldots, 8$ we obtain 250 resamples, by sampling with replacement between the 1,882 phrases. For each resample and for each pair of consecutive trees (τ_i, τ_{i+1}), we compute the difference between the logarithm of the likelihoods and divide this quantity by n_j. The corresponding boxplots for each value of n_j and for the trees with 8, 11, 13, 16 and 17 leaves is presented in Figure 8. Note that we can clearly see a change of behavior in the boxplots when considering trees bigger or equal the real tree. More specifically, for the boxplots corresponding either to the pair of trees with 8 and 11 leaves or the pair of trees with 11 and 13 leaves, the median seems to fluctuate around a constant value. In opposition, for the boxplots corresponding either to the pair of trees with 13 and 16 leaves or the pair of trees with 16 and 17 leaves, the medians seem approaches zero when the sample size increases.



w	$p(0\|w)$	$p(1\|w)$	$p(2\|w)$	$p(3\|w)$	$p(4\|w)$
000	0.27	0.73	0.00	0.00	0.00
100	0.00	0.00	0.67	0.25	0.08
200	0.36	0.64	0.00	0.00	0.00
300	0.00	0.00	0.70	0.20	0.10
010	0.05	0.00	0.67	0.19	0.09
210	0.08	0.00	0.63	0.22	0.07
20	0.45	0.55	0.00	0.00	0.00
30	0.05	0.00	0.64	0.27	0.04
001	0.61	0.00	0.28	0.07	0.04
201	0.72	0.00	0.19	0.07	0.02
21	0.72	0.00	0.19	0.07	0.02
02	0.59	0.41	0.00	0.00	0.00
012	0.58	0.42	0.00	0.00	0.00
212	0.53	0.47	0.00	0.00	0.00
32	0.50	0.50	0.00	0.00	0.00
42	0.52	0.48	0.00	0.00	0.00
3	0.69	0.00	0.19	0.12	0.00
4	0.00	0.00	0.65	0.35	0.00

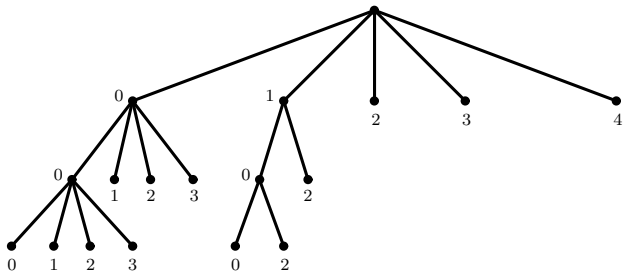
Figure 5: Probabilistic context trees for European Portuguese.

7 Final discussion

In this paper we introduce the smallest maximizer criterion to estimate the context tree of a chain with memory of variable length from a finite sample. The criterion selects a tree in the class of champion trees. This class coincides with the subset of trees obtained by varying the penalizing constant in the BIC criterion. For this reason, the smallest maximizer criterion actually suggests a tuning procedure for the BIC context tree selection. Therefore, the present paper can be interpreted as an effort to solve the most important problem of constant-free model selection in the case of probabilistic context tree models.

To our knowledge Bühlmann (2000) was the first to address the problem of how to tune a context tree estimator, in the case of the algorithm Context. This paper proposes the following tuning procedure. First use the algorithm Context with different values of the threshold to obtain a sequence of candidate trees. For each one of these candidate trees estimate a global risk function, as for example the Final Prediction Error (FPE) or the Kullback-Leibler Information (KLI), by using a parametric bootstrap approach. Then choose as cut-off parameter the one providing the tree with smallest estimated risk.

In the above mentioned paper there is no proof that the sequence of nested trees obtained by the pruning procedure using the algorithm Context will contain eventually almost surely the tree generating the sample, which in our case is given in Theorem 3.4. It also misses



w	$p(0\|w)$	$p(1\|w)$	$p(2\|w)$	$p(3\|w)$	$p(4\|w)$
000	0.29	0.71	0.00	0.00	0.00
100	0.00	0.00	0.67	0.21	0.12
200	0.40	0.60	0.00	0.00	0.00
300	0.00	0.00	0. 67	0.22	0.11
10	0.07	0.00	0.65	0.21	0.07
20	0.45	0.55	0.00	0.00	0.00
30	0.07	0.00	0.64	0.25	0.04
001	0.62	0.00	0.27	0.08	0.03
201	0.72	0.00	0.19	0.07	0.02
21	0.73	0.00	0.18	0.08	0.01
2	0.60	0.40	0.00	0.00	0.00
3	0.69	0.00	0.21	0.10	0.00
4	0.00	0.00	0.66	0.34	0.00

Figure 6: Context tree and transition probabilities over the alphabet $A = \{0, 1, 2, 3, 4\}$.

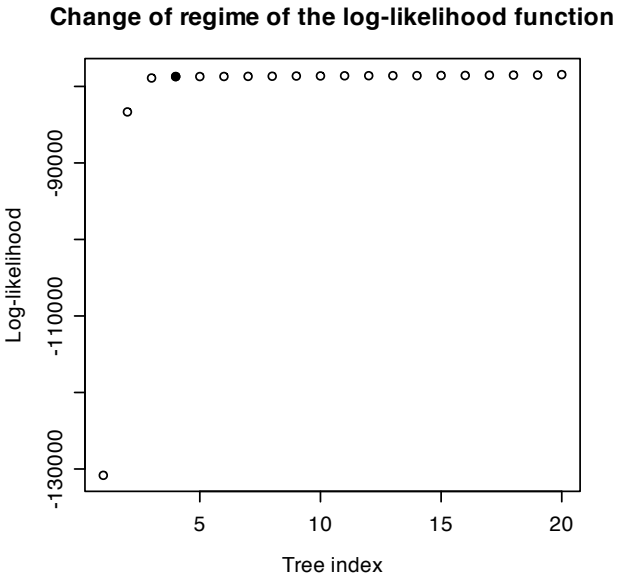


Figure 7: Change of regime of the log-likelihood function. The true tree corresponds to the filled point.

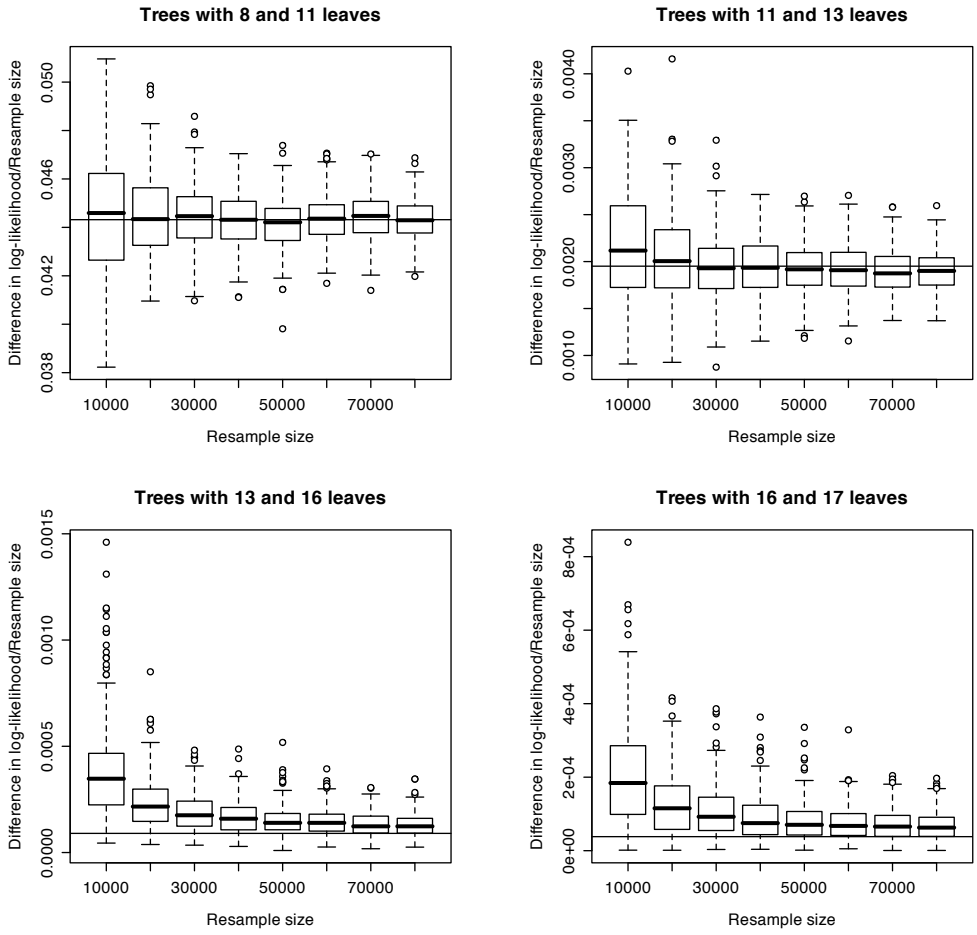


Figure 8: Simulation results.

the crucial point of the change of regime in the set of champion trees, which is given in our Theorem 3.5.

The change of regime was not missed by the more recent paper of Dalevi and Dubhashi (2005). They extend to chains with memory of variable length the order estimator introduced in Peres and Shields (2005). They suggest without any rigorous proof that at the correct order there exists a sharp transition that can be identified from a finite sample. Then they applied the criterion to the identification of sequence similarity in DNA. Our main contribution with respect to this paper is the rigorous proof of Theorem 3.6.

Acknowledgments

We thank D. Brillinger, F. Cribari, R. Dias, D. Duarte, J. Goldsmith, C. Peixoto, C. Robert and D. Takahashi for discussions, comments and bibliographic suggestions. This work is part of PRONEX/FAPESP's project *Stochastic behavior, critical phenomena and rhythmic pattern identification in natural languages* (grant number 03/09930-9) and CNRS-FAPESP project *Probabilistic phonology of rhythm.* and CNPq's project *Rhythmic patterns, prosodic*

domains and probabilistic modeling in Portuguese Corpora (grant number 485999/2007-2). AG, CG and NLG are partially supported by a CNPq fellowship (grants 308656/2005-9, 303421/2004-5 and 301530/2007-6, respectively).

Appendix 1 - Mathematical proofs

Proof of Proposition 4.3 Let us first show that for any $c > 0$, $\hat{\tau}_{\text{BIC}}(X_1^n, c)$ belongs to $\mathcal{C}_n$. Recall that $\mathcal{T}_n = \cup_{g \in \mathcal{G}_n} \mathcal{T}_n^g$. Therefore, for any $c > 0$

$$\begin{aligned}\arg\max_{\tau \in \mathcal{T}_n}\{\log L_\tau(X_1^n) - c \cdot \text{df}(\tau) \cdot \log n\} &= \arg\max_{g \in \mathcal{G}_n} \max_{\tau \in \mathcal{T}_n^g}\{\log L_\tau(X_1^n) - c \cdot \text{df}(\tau) \cdot \log n\} \\ &= \arg\max_{g \in \mathcal{G}_n}\{\log L_{\tau_n^g}(X_1^n) - c \cdot g \cdot \log n\}.\end{aligned}$$

Since $\mathcal{G}_n$ is finite, for each $c > 0$ the maximum in the above equation is reached. Since different champion trees have different likelihood there exists only one champion tree corresponding to each constant $c > 0$.

Now we have to prove that for any $\tau_n^g \in \mathcal{C}_n$, there exists a positive constant $c = cg$ such that $\tau_n^g = \hat{\tau}_{\text{BIC}}(X_1^n; c)$. By definition, for any two champion trees $\tau_n^{g'}$ and $\tau_n^{(g'')}$ belonging to $\mathcal{C}_n$, with $g' < g''$, we have

$$\log L_{\tau_n^{g'}}(X_1^n) < \log L_{\tau_n^{(g'')}}(X_1^n).$$

Therefore, the rate

$$\frac{\log L_{\tau_n^{g'}}(X_1^n) - \log L_{\tau_n^{(g'')}}(X_1^n)}{g' - g''}$$

is always positive. The result follows by choosing c as

$$c = \min\left\{\frac{\log L_{\tau_n^{g'}}(X_1^n) - \log L_{\tau_n^g}(X_1^n)}{g' - g}, g' \in \mathcal{G}_n \setminus \{g\}\right\}.$$

This concludes the proof of the proposition.

The tools to prove Theorems 3.4 and 3.5 are borrowed from Csiszár and Talata (2006).

Proof of Theorem 3.4. First recall that the BIC context tree estimator is strongly consistent for any constant $c > 0$. Therefore, since the set $\mathcal{C}$ is countable, it follows that eventually almost surely $\tau^* \in \mathcal{C}_n$ as $n \to \infty$.

The fact that the champion trees are ordered by $\prec$ follows immediately from the following lemma and Proposition 4.3.

Lemma 8.1. *Let $0 < c_1 < c_2$ be arbitrary positive constants. Then*

$$\hat{\tau}_{\text{BIC}}(X_1^n; c_1) \succeq \hat{\tau}_{\text{BIC}}(X_1^n; c_2).$$

Proof. For a string w with $\ell(w) \le d(n)$ define

$$L_w(X_1^n) = \prod_{a\in A} \hat{p}_n(a|w)^{N_n(wa)}$$

and $\mathrm{df}(w) = \sum_{a\in A} \chi(wa)$. Then, for any constant $c > 0$ define recursively the value

$$V_w^c(X_1^n) = \begin{cases} \max\{n^{-c\cdot \mathrm{df}(w)} L_w(X_1^n),\ \prod_{a\in A} V_{aw}^c(X_1^n)\}, & \text{if } 0 \le \ell(w) < d(n), \\ n^{-c\cdot \mathrm{df}(w)} L_w(X_1^n), & \text{if } \ell(w) = d(n) \end{cases}$$

and the indicator

$$\delta_w^c(X_1^n) = \begin{cases} 1, & \text{if } 0 \le \ell(w) < d(n) \text{ and } \prod_{a\in A} V_{aw}^c(X_1^n) > n^{-c\cdot \mathrm{df}(w)} L_w(X_1^n), \\ 0, & \text{if } 0 \le \ell(w) < d(n) \text{ and } \prod_{a\in A} V_{aw}^c(X_1^n) \le n^{-c\cdot \mathrm{df}(w)} L_w(X_1^n), \\ 0, & \text{if } \ell(w) = d(n). \end{cases}$$

Now, for any finite string w, with $\ell(w) \le d(n)$ and for any tree $\tau \in \mathcal{T}_n$, we define the irreducible tree τ_w as the set of branches in τ which have w as a suffix, that is

$$\tau_w = \{u \in \tau \colon w \preceq u\}.$$

Let $\mathcal{T}_w(X_1^n)$ be the set of all trees defined in this way, that is

$$\mathcal{T}_w(X_1^n) = \{\tau_w \colon \tau \in \mathcal{T}_n\}.$$

If w is a sequence such that $\delta_w^c(X_1^n) = 1$ we define the maximizing tree assigned to the sequence w as the tree $\tau_w^M(X_1^n) \in \mathcal{T}_w(X_1^m)$ given by

$$\tau_w^M(X_1^n) = \{u \colon \delta_u^c(X_1^n) = 0,\ \delta_v^c(X_1^n) = 1 \text{ for all } w \preceq v \prec u\}.$$

If w is a sequence such that $\delta_w^c(X_1^n) = 0$, we define the maximizing tree assigned to the sequence w as the tree $\tau_w^M(X_1^n) \in \mathcal{T}_w(X_1^m)$ given by

$$\tau_w^M(X_1^n) = \{w\}.$$

Csiszár and Talata (2006) proved that

$$V_w^c(X_1^n) = \max_{\tau\in T_w(X_1^n)} \prod_{u\in\tau} n^{-c\cdot \mathrm{df}(u)} L_u(X_1^n) = \prod_{u\in\tau_w^M(X_1^n)} n^{-c\cdot \mathrm{df}(u)} L_u(X_1^n). \tag{8.2}$$

Denote by $\tau^1 = \hat{\tau}_{\mathrm{BIC}}(X_1^n; c_1)$ and $\tau^2 = \hat{\tau}_{\mathrm{BIC}}(X_1^n; c_2)$. Suppose that it is not true that $\tau^1 \succeq \tau^2$. Then there exists a sequence $w \in \tau^1$ and $w' \in \tau^2$ such that w is a proper suffix of w'. This implies that $\tau_w^2 \neq \emptyset$. Since τ^2 is irreducible we have that $|\tau_w^2| \ge 2$. Then, using the definition of maximizing tree we obtain

$$\begin{aligned} \log L_w(X_1^n) &\ge \sum_{w'\in\tau_w^2} \log L_{w'}(X_1^n) + c_1(\mathrm{df}(w) - \sum_{w'\in\tau_w^2} \mathrm{df}(w')) \log n \\ &\ge \sum_{w'\in\tau_w^2} \log L_{w'}(X_1^n) + c_2(\mathrm{df}(w) - \sum_{w'\in\tau_w^2} \mathrm{df}(w')) \log n \\ &> \log L_w(X_1^n), \end{aligned}$$

which is a contradiction. The first inequality follows from the assumption that $\tau^1 = \hat{\tau}_{\text{BIC}}(X_1^n; c_1)$ and the second equality in (8.2). To derive the second inequality we use the fact that $0 < c_1 < c_2$ and $\text{df}(w) - \sum_{w' \in \tau_w^2} \text{df}(w') < 0$. Finally, the last inequality leading to the contradiction follows from $\tau^2 = \hat{\tau}_{\text{BIC}}(X_1^n; c_2)$ and again the second equality in (8.2). We conclude that $\tau^1 \succeq \tau^2$.

Proof of Theorem 3.5 To prove (1) let $\tau \in \mathcal{C}_n$ be such that $\tau \prec \tau^*$. Then

$$\begin{aligned}
&\log L_\tau(X_1^n) - \log L_{\tau^*}(X_1^n) \\
&\qquad = \sum_{w' \in \tau, a \in A} N_n(w'a) \log \hat{p}_n(a|w') - \sum_{w \in \tau^*, a \in A} N_n(wa) \log \hat{p}_n(a|w) \\
&\qquad = \sum_{w' \in \tau} \sum_{w \in \tau^*, w \succ w'} \sum_{a \in A} N_n(wa) \log \frac{\hat{p}_n(a|w')}{\hat{p}_n(a|w)}.
\end{aligned}$$

Dividing by n and using Jensen's inequality in the right hand side we have that

$$\sum_{w' \in \tau} \sum_{w \in \tau^*, w \succ w'} \sum_{a \in A} \frac{N_n(wa)}{n} \log \frac{\hat{p}_n(a|w')}{\hat{p}_n(a|w)} \longrightarrow \sum_{w' \in \tau'} \sum_{w \in \tau^*, w \succ w'} \sum_{a \in A} p^*(wa) \log \frac{p^*(a|w')}{p^*(a|w)} \; < \; 0,$$

as n goes to $+\infty$ (by the minimality of τ^*). Then, for a sufficiently large n there exists a constant $c(\tau^*, \tau) > 0$ such that

$$\log L_{\tau^*}(X_1^n) - \log L_\tau(X_1^n) \geq c(\tau^*, \tau) n.$$

To prove (2) we have that

$$\begin{aligned}
&\log L_{\tau'}(X_1^n) - \log L_\tau(X_1^n) \\
&\qquad = \sum_{w' \in \tau', a \in A} N_n(w'a) \log \hat{p}_n(a|w') - \sum_{w \in \tau, a \in A} N_n(wa) \log \hat{p}_n(a|w) \\
&\qquad \leq \sum_{w' \in \tau', a \in A} N_n(w'a) \log \hat{p}_n(a|w') - \sum_{w \in \tau, a \in A} N_n(wa) \log p^*(a|w) \\
&\qquad = \sum_{w \in \tau} \sum_{w' \in \tau', w' \succ w} \sum_{a \in A} N_n(w'a) \log \frac{\hat{p}_n(a|w')}{p^*(a|w)} \\
&\qquad = \sum_{w \in \tau} \sum_{w' \in \tau', w' \succ w} N_n(w'\cdot) D(\hat{p}_n(\cdot|w') \| p^*(\cdot|w)).
\end{aligned}$$

By Lemmas 6.2 and 6.3 in Csiszár and Talata (2006) we have that, if n is sufficiently large, we can bound above the last term by

$$\begin{aligned}
&\sum_{w \in \tau} \sum_{w' \in \tau', w' \succ w} N_n(w'\cdot) \sum_{a \in A} \frac{[\hat{p}_n(a|w') - p^*(a|w)]^2}{p^*(a|w)} \\
&\qquad\qquad \leq \sum_{w \in \tau} \sum_{w' \in \tau', w' \succ w} N_n(w'\cdot) \frac{1}{p^*_{\min}} |A| \frac{\delta \log n}{N_n(w'\cdot)},
\end{aligned}$$

where $p^*_{\min} = \min_{w \in \tau, a \in A} \{p^*(a|w) \colon p^*(a|w) > 0\}$. This concludes the proof of Theorem 3.5.

Proof of Theorem 3.6 It follows directly from Theorems 3.4 and 3.5.

Appendix 2 - Description of the encoded samples

The newspapers articles of the sample were selected in the following way. We first randomly selected 20 editions for each newspaper for each year. Inside each edition we discarded all the texts with less than 1000 words as well as some type of articles (interviews, synopsis, transcriptions of laws and collected works) which are unsuitable for our purposes. From the remaining articles we randomly selected one article for each previously selected edition.

Before encoding each one of the selected texts, they were submitted to a linguistically oriented cleaning procedure. Hyphenated compound words were rewritten as two separate words, except when one of the components is unstressed. Suspension points, question marks and exclamation points were replaced by periods. Dates and special symbols like "%" were spelled out as words. All parentheses were removed.

To use the smallest maximizer criterion we need to compute the number of degrees of freedom of each candidate context tree. To do this we must take into account the linguistic restrictions on the symbolic chain obtained after encoding. The restrictions are the following.

1. Due to Portuguese morphological constraints, a stressed syllable (encoded by 1 or 3) can be immediately followed by at most three unstressed syllables (encoded by 0) .
2. Since by definition any prosodic word must contain one and only one stressed syllable (encoded by 1 or 3), after a symbol 3 no symbol 1 is allowed, before a symbol 2 (non stressed syllable starting a prosodic word) appears.
3. By the same reason, after a symbol 2 no symbols 2 or 3 are allowed before a symbol 1 appears.
4. As sentences are formed by the concatenation of prosodic words, the only symbols allowed after 4 (end of sentence) are the symbols 2 or 3 (beginning of prosodic word).

References

[1] P. Bühlmann. Model selection for variable length Markov chains and tuning the context algorithm. *Ann. Inst. Statist. Math.*, 52(2):287–315, 2000.

[2] P. Bühlmann and A. J. Wyner. Variable length Markov chains. *Ann. Statist.*, 27:480–513, 1999.

[3] J. Busch, P. Ferrari, A.G. Flesia, R. Fraiman, S. Grynberg, and F. Leonardi. Testing statistical hypothesis on random trees and applications to the protein classification problem. *Ann. Appl. Stat.*, 3(2):542–563, 2009.

[4] N. Chomsky. *Knowledge of language: its nature, origins, and use.* New York: Praeger, 1985.

[5] I. Csiszár and Z. Talata. Context tree estimation for not necessarily finite memory processes, via BIC and MDL. *IEEE Trans. Inform. Theory*, 52(3):1007–1016, 2006.

[6] J.A. Cuesta-Albertos, R. Fraiman, A. Galves, J. Garcia, and M. Svarc. Identifying rhythmic classes of languages using their sonority: a Kolmogorov-Smirnov approach. *Journal of Applied Statistics*, 34(6):749–761, 2007.

[7] D. Dalevi and D. Dubhashi. The Peres-Shields order estimator for fixed and variable length markov models with applications to DNA sequence similarity. *Algorithms in Bioinformatics*, pages 291–302, 2005.

[8] S. Frota and M. Vigário. On the correlates of rhythm distinctions: the European/Brazilian Portuguese case. *Probus*, 13:247–275, 2001.

[9] A. Galves and E. Löcherbach. Stochastic chains with memory of variable length. *Festschrift for Jorma Rissanen, Grünwald et al. (eds)*, TICSP Series 38:117–133, 2008.

[10] A. Galves, V. Maume-Deschamps, and B. Schmitt. Exponential inequalities for VLMC empirical trees. *ESAIM Prob. Stat.*, 12:219–229, 2008.

[11] A. Garivier. *Modèles contextuels et alphabets infinis en théorie de l'information*. PhD thesis, Université de Paris Sud, 2006.

[12] Y. Peres and P. Shields. Two new Markov order estimators, 2005. arXiv:math/0506080v1 [math.ST].

[13] F. Ramus. Acoustic correlates of linguistic rhythm: perspectives. In *Speech Prosody 2002*, pages 115–120, 2002.

[14] I. S. Révah. L'évolution de la prononciation au Portugal et au Brésil du XVI^{e} siècle à nos jours. In *Anais do primeiro Congresso Brasileiro da língua falada no teatro*, 1958.

[15] J. Rissanen. A universal data compression system. *IEEE Trans. Inform. Theory*, 29(5):656–664, 1983.

[16] D. Ron, Y. Singer, and N. Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 25(2-3):117–149, 1996.

[17] F. Sândalo, M. B. Abaurre, A. Mandel, and C. Galves. Secondary stress in two varieties of portuguese and the sotaq optimality based computer program. *Probus*, 18:97–125, 2006.

[18] D. Santos and P. Rocha. *The key to the first CLEF in Portuguese: Topics, questions and answers in CHAVE*, pages 821–832. Lecture Notes in Computer Science. Berlin/Heidelberg:Springer, 2005.

[19] M. Vigário. *The prosodic word in European Portuguese*. Berlim/New York: Mouton Degruyter, 2003.

[20] F.M.J. Willems, Y.M. Shtarkov, and T.J. Tjalkens. The context-tree weighting method: Basic properties. *IEEE Trans. Inf. Theory*, 41(3):653–664, 1995.