Model Selection via Information Criteria for Tree Models and Markov Random Fields

By

Zsolt Talata

Ph.D. Dissertation

Thesis advisor: Professor Imre Csiszár
Rényi Institute of Mathematics
Hungarian Academy of Sciences

Institute of Mathematics
Budapest University of Technology and Economics
Budapest, Hungary

2004

Contents

Preface

The dissertation deals with *model selection problems*. Chapter 1 is a survey of these statistical problems. They can be formulated as follows. Let a stochastic process be given that we would like to model. Further, let a family of model classes be given, each class determined by a structure parameter. Each model in a class is described by a parameter vector from a subset of an Euclidean space whose dimension depends on the structure parameter. Suppose that based on a realization of the process, called statistical *sample*, we can estimate the parameter vector provided the structure parameter is known. The task is estimation of the latter. Examples of model classes are autoregressive (AR) processes, ARMA processes, Markov chains, tree models, Markov random fields.

The dissertation treats the model selection problem using the concept of *information criterion*. An information criterion assigns a real number to each hypothetical model class, the structure parameter is estimated by minimizing this criterion. The mostly used information criteria are the *Bayesian Information Criterion (BIC)* and the *Minimum Description Length (MDL)*. The BIC consists of two terms. The first one is the negative logarithm of the maximum likelihood, this measures the goodness of fit of the sample to the model class. The second term is the half number of free parameters in the model class times the logarithm of the sample size, this penalizes too complex models. The MDL is based on a code of the sample tailored to the model class, and on a code of the structure parameter; the sum of codelengths of these codes gives the criterion.

An estimator of the structure parameter is said to be *strongly consistent* if, with probability 1, it equals the true structure parameter when the sample size is sufficiently large. It has been known for various model classes that the BIC and MDL estimators are strongly consistent, mostly under the assumption that the true model belongs to a known finite set of model classes. It has been proved recently that in the case of Markov chains, the latter assumption can be dropped; for BIC, there is no need for any bound on the hypothetical order at all, for MDL one can use a bound that grows with the sample size. The dissertation, motivated by these results, presents new results in two areas.

In Chapter 2, the concept of *context tree*, usually defined for finite memory processes, is extended to arbitrary stationary ergodic processes (with finite alphabet). These context trees are not necessarily complete, and may be of infinite depth. The BIC and MDL principles are shown to provide strongly consistent estimators of the context tree; here the depth of the hypothetical context trees is allowed to grow with the sample size n as $o(\log n)$, hence there is no need for a prior bound on the true depth. Moreover, algorithms are provided to compute these estimators in $O(n)$ time, and to compute them on-line for all $i \leq n$ in $o(n \log n)$ time. In the MDL case the algorithm is a modification of a known method. It is important that this method can also be extended for the BIC case, because previously the BIC estimator of the context tree was believed to be computationally infeasible.

In Chapter 3, for *Markov random fields* on $\mathbb{Z}^d$ with finite state space, the statistical estimation of the basic neighborhood is addressed. The basic neighborhood is the smallest region that determines the conditional distribution at a site on the condition that the values at all other sites are given. Here the samples are observations of a realization of the field on increasing finite regions. The BIC and MDL estimators are unsuitable for this problem, but a modification of BIC, replacing likelihood by *pseudo-likelihood*, is proved to provide a strongly consistent estimator. The size of the hypothetical basic neighborhood may extend with the sample size, thus no prior bound on the size of the true basic neighborhood is required.

Each part of the dissertation is published. The three chapters correspond to three papers. Essentially, Chapters 1, 2 and 3 are the papers (Talata, 2004), (Csiszár and Talata, 2004b) and (Csiszár and Talata, 2004a), respectively. The almost only difference between the papers and the dissertation is that the references are merged into the end of the dissertation.

The referees' report on the dissertation and the minutes of the defense of thesis will be available at Dean's Office, Faculty of Science, Budapest University of Technology and Economics.

I thank Professor Imre Csiszár for being my thesis advisor. I am glad to work with him, I have been learning a lot from him.

Finally, I declare the following.

I, the undersigned Zsolt Talata, state that I produced this dissertation myself and I used only the indicated sources for it. Each part, which I adopted literally or with the same content but rewritten, is referred unambiguously, with indication the source.

This declaration in Hungarian:

Alulírott Talata Zsolt kijelentem, hogy ezt a doktori értekezést magam készítettem és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelűen, a forrás megadásával megjelöltem.

Budapest, December 3, 2004.

Zsolt Talata

Chapter 1

Introduction

1.1 The model selection problem

Let a stochastic process $\{ X_t, t \in T \}$ be given, where each X_t is a random variable with values $a \in A$, and T is an index set. The joint distribution of the random variables X_t, $t \in T$ will be referred to as the distribution of the process and will be denoted by Q. A *model* of the process determines a hypothetical distribution of the process or a collection of hypothetical distributions. Typically, a model is determined by a structure parameter k with values in some set $\mathcal{K}$, and by a parameter vector $\theta_k \in \Theta_k \subset \mathbb{R}^{d_k}$; this model is denoted by M_{θ_k}. Given the feasible models of the process, they can be arranged into model classes according to the structure parameter: $\mathcal{M}_k = \{ M_{\theta_k}, \theta_k \in \Theta_k \subset \mathbb{R}^{d_k} \}$. Statistical inference about the process is drawn based on a realization $\{ x_t, t \in T \}$ of the process observed in the range $R_n \subset T$, where R_n extends with n. Thus the n'th *sample* is $x(n) = \{ x_t, t \in R_n \}$. Some typical examples for processes and their models are listed below.

In the case of *density function estimation*, $T = \mathbb{N}$ and the random variables X_t, $t \in \mathbb{N}$ are independent and identically distributed (i.i.d.) with density function f_{θ_k}. The n'th sample is $\{ x_i, i = 1, \ldots, n \}$.

The *polynomial fitting* involves $T \subseteq \mathbb{R}$, where T is a countable set, $A = \mathbb{R}$, and the model

$$X_t = \theta_k[0] + \theta_k[1]\, t + \theta_k[2]\, t^2 + \cdots + \theta_k[k-1]\, t^{k-1} + Z_t,$$

where Z_t, $t \in T$ are independent random variables with normal distribution, zero mean, unknown common variance, and $\theta_k[i]$ is the i'th component of the k-dimensional parameter vector θ_k. Here the structure parameter $k \in \mathbb{N}$ is the degree of the polynomial $\theta_k[0] + \theta_k[1]\, t + \theta_k[2]\, t^2 + \cdots + \theta_k[k-1]\, t^{k-1}$ plus 1, and the n'th sample is $\{ x_t, t \in \{t_1, \ldots, t_n\} \subset T \}$.

The process with $T = \mathbb{N}$, $A = \mathbb{R}$ is an *autoregressive (AR) process* of order k if

$$X_t = \sum_{i=1}^{k} a_i X_{t-i} + Z_t,$$

where Z_t, $t \in \mathbb{N}$ are independent random variables with normal distribution, zero mean, unknown common variance, and $a_i \in \mathbb{R}$, $i = 1, \ldots, k$ form the parameter vector θ_k. Here the structure parameter $k \in \mathbb{N}$ is the number of coefficients a_i, and the n'th sample is $\{ x_i, i = 1, \ldots, n \}$.

The *autoregressive moving average (ARMA) process* is similar to the AR process. In this case we have

$$X_t = \sum_{i=1}^{p} a_i X_{t-i} + Z_t + \sum_{j=1}^{q} b_j Z_{t-j}.$$

The parameter vector is $\theta_k = \{ a_1, \ldots, a_p, b_1, \ldots, b_q \} \in \mathbb{R}^{p+q}$, and the structure parameter k has two components: $k = (p, q) \in \mathbb{N}^2$.

The process with $T = \mathbb{N}$, $|A| < \infty$ is a *Markov chain* of order k if

$$Q(X_1^n = x_1^n) = Q(X_1^k = x_1^k) \prod_{i=k+1}^{n} Q(x_i | \, x_{i-k}^{i-1}), \quad n \geq k, \, x_1^n \in A^n, \tag{1.1}$$

with suitable transition probabilities $Q(\cdot \mid \cdot)$. Here x_i^j denotes the sequence $x_i, x_{i+1}, \ldots, x_j$. Since for each $a_1^k \in A^k$ the vector $\{ Q(a | \, a_1^k), \, a \in A \}$ gives a probability distribution on A, the parameter vector $\theta_k \in \mathbb{R}^{d_k}$ consists of $d_k = (|A| - 1) \, |A|^k$ transition probabilities $Q(a | \, a_1^k)$, $a \in A^*$, $a_1^k \in A^k$, where $|A^*| = |A| - 1$. Here the structure parameter $k \in \mathbb{N}$ is the length of the sequence that the transitional probabilities depend on in their second argument. The n'th sample is $\{ x_i, i = 1, \ldots, n \}$.

The AR and ARMA processes, and Markov chains are examples for the case when the model does not determine a unique hypothetical distribution of the process. In particular, for AR processes or Markov chains of order k the model determines only a hypothetical conditional distribution for $X_{k+1}, X_{k+2}, \ldots$ given $X_1, \ldots, X_k$.

The set $\mathcal{K}$ of feasible structure parameters k is an ordered or partially ordered set with respect to the inclusion of the model classes $\mathcal{M}_k$. When the model M_{θ_k} with structure parameter k corresponds to the true distribution Q of the process, a more complex model with (in the above sense) greater structure parameter k' may also correspond to the distribution Q with a suitable parameter vector $\theta_{k'}$. For example, any AR process or Markov chain of order k is also of order k', for each $k' > k$. We mean by the *true model* M_{θ_0} the minimal model among those

that correspond to the true distribution Q, that is, for which there exists no other model with the same property that has a smaller structure parameter in the above sense. The structure parameter of this true model will be denoted by k_0.

The *model selection problem* consists in estimating the true structure parameter k_0 based on the statistical observation $x(n)$ of the process.

The term *underestimation* refers to the case when a smaller structure parameter k is selected than the true one k_0. In such a case $\theta_0 \notin \Theta_k$, hence the true model cannot be estimated accurately; the estimation of the parameter vector will involve bias.

The term *overestimation* refers to the case when a greater structure parameter k is selected than the true one k_0. In this case $M_{\theta_0} \in \mathcal{M}_{k_0} \subset \mathcal{M}_k$, thus $M_{\theta_0} = M_{\theta_k}$ for some $\theta_k \in \Theta_k$, but θ_k has more components than θ_0, hence it is more difficult to estimate the true setting; the estimation of the parameter vector will have larger variance.

The dissertation treats the model selection problem using the concept of information criterion. An *information criterion (IC)* based on the sample $x(n)$ assigns a real value to each model class: $IC : \mathcal{K} \times \{x(n)\} \to \mathbb{R}$, and the estimator of k_0 equals the structure parameter with the minimum value of the criterion:

$$\hat{k}(x(n)) = \arg\min_{k \in \mathcal{K}} IC_k(x(n)).$$

The next sections give an overview about information criteria.

1.2 Historical review

The model selection problem can be regarded as multiple hypothesis testing, and the *likelihood ratio test* procedure of Neyman and Pearson (1928) can be used. Anderson worked out this procedure for polynomial fitting (Anderson, 1962) and for AR processes (Anderson, 1963). These procedures are sequences of tests taking the hypothetical orders successively, starting at the highest one. The main disadvantage of these procedures is the subjective choice of the significance levels of the tests for all hypothetical model orders.

Mallows (1964, 1973) introduced, for selecting the true variables of linear models, a method similar to the information criteria. Consider the *linear model*

$$X_t = \sum_{i=1}^{K} a_i u_{it} + a_0 + Z_t, \quad t \in \mathbb{Z},$$

where a_i, $i = 1, \ldots, K$ are the parameters of the model, u_{it}, $i = 1, \ldots, K$ are (non-random) independent variables whose values are given at $t = 1, \ldots, n$, and Z_t's are independent random variables with zero mean and unknown common variance σ^2. Given the sample $x(n) = \{ x_t, t = 1, \ldots, n \}$, the problem is to estimate the set $\{ u_{i_1}, \ldots, u_{i_k} \}$ of variables that X_t effectively depends on, that is, $a_{i_l} \neq 0$ for $i_l \in \{ i_1, \ldots, i_k \}$ and $a_{i_l} = 0$ otherwise.

Mallows assigned to each hypothetical index set $P = \{ i_1, \ldots, i_k \}$ the value

$$C_P = \frac{1}{\hat{\sigma}^2} \mathrm{RSS}_P - n + 2|P|,$$

where RSS_P is the residual sum of squares according to P:

$$\mathrm{RSS}_P = \min_{a_{i_l},\, i_l \in P} \sum_{t=1}^{n} \left(x_t - \sum_{a_{i_l},\, i_l \in P} a_{i_l} u_{i_l t} - a_0 \right)^2,$$

moreover $\hat{\sigma}^2$ is a suitable estimate of σ^2, e.g., $\hat{\sigma}^2 = \mathrm{RSS}_{\{1,\ldots,k\}}/(n-k)$. The estimator is the index set P with minimum C_P. It can be shown that the expected value of C_P is equal to $|P|$ when P is the true index set, and it is greater otherwise.

For stationary processes, Davisson (1965) analyzed the mean square prediction error of the AR model of order k, when the coefficients of the model are determined based on the past n observations $x_1, \ldots, x_n$ and this model is applied to predict the next observation. Namely, for the predictor $\hat{X}_n(k) = \sum_{i=1}^{k} \hat{a}_i X_{n-i}$ with coefficients which minimize the mean square prediction error, that is,

$$\{ \hat{a}_1, \ldots, \hat{a}_k \} = \arg \min_{\{a_1, \ldots, a_k\}} \sum_{t=0}^{n-1} \left(x_t - \sum_{i=i}^{k} a_i x_{t-i} \right)^2,$$

he obtained

$$\mathbb{E} \left(X_n - \hat{X}_n(k) \right)^2 = \sigma^2(k) \left(1 + \frac{k}{n} \right) + o\left(\frac{1}{n} \right),$$

where $\sigma^2(k)$ is the asymptotic mean square error. Moreover, he suggested using the main term of the above expression to estimate the true order, via minimizing it over the candidate orders. Of course, this requires the estimation of $\sigma^2(k)$.

Akaike (1970) arrived at the same result, and he overcame the problem of estimating $\sigma^2(k)$ by a suitable spectral estimation method. He defined a criterion called *final prediction error* as

$$\mathrm{FPE}_k(x_1^n) = \frac{n+k}{n-k} \left(\hat{C}_0 - \hat{a}_1 \hat{C}_1 - \cdots - \hat{a}_k \hat{C}_k \right),$$

where $\hat{C}_i = (1/n) \sum_{t=1}^{n-1} x_{t+l} x_t$, $i = 0, \ldots, k$ are the correlation coefficients, and $\hat{a}_i$, $i = 1, \ldots, k$ are the model coefficients which minimize the least square prediction

error, as above. The latter values can be calculated from the $\hat{C}_i$'s, solving the Yule – Walker equation. The order estimator is

$$\hat{k}(x_1^n) = \arg \min_{0 \le k \le K} \mathrm{FPE}_k(x_1^n).$$

The only subjective element in this procedure is the determination of the upper bound K of candidate orders. Akaike also showed that this estimator overestimates the true order asymptotically with positive probability, that is,

$$\liminf_{n \to \infty} Q\left(\hat{k}(x_1^n) > k_0\right) > 0.$$

Akaike (1972) introduced a general concept for solving the model selection problem. Assume that each model M_{θ_k} specifies a unique distribution P_{θ_k} of the process, and let $P_{\theta_k}^{(n)}$ denote its marginal equal to the distribution of the sample $x(n)$. The *Kullback – Leibler information divergence* between $P_{\theta_k}^{(n)}$ and $P_{\theta_0}^{(n)}$ is

$$\mathrm{D}\left(P_{\theta_0}^{(n)} \middle\| P_{\theta_k}^{(n)}\right) = \int f_{\theta_0}^{(n)}(x(n)) \log \frac{f_{\theta_0}^{(n)}(x(n))}{f_{\theta_k}^{(n)}(x(n))} \, \lambda(\mathrm{d}x(n)),$$

where $f_{\theta_k}^{(n)}$ denotes the density of $P_{\theta_k}^{(n)}$ with respect to a dominating measure λ (typically, λ is either the Lebesgue measure or, in the discrete case, the counting measure). Logarithms are to the base e. Akaike aimed at minimization of this quantity for estimating the true parameter vector θ_0 and the true structure parameter k_0. He found that this minimizer can be approximated by taking the maximum likelihood estimator $\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} f_{\theta_k}^{(n)}(x(n))$ in each candidate model class, and then selecting the model class whose structure parameter minimizes the value

$$\mathrm{AIC}_k(x(n)) = -\log f_{\hat{\theta}_k}^{(n)}(x(n)) + \dim \Theta_k.$$

When the models do not determine uniquely the distribution of the process, we can define the AIC similarly, with suitably defined $f_{\theta_k}^{(n)}$. For example, in the case of AR process of order k we can prescribe $X_1, \ldots, X_k$ to be 0, or to have the marginal distribution of the stationary distribution of the process. This specifies a unique joint distribution corresponding to the model, and we can take its density as $f_{\theta_k}^{(n)}$. Note that suitable restriction on the parameter set Θ_k can guarantee the existence of the stationary distribution. In the case of Markov chains of order k, we can either proceed similarly, or we can define $f_{\theta_k}^{(n)}$ as the right hand side of (1.1) dropping the factor $Q(X_1^k = x_1^k)$.

This model selection procedure has a clear interpretation. The first term of the information criterion is the negative logarithm of the maximum likelihood.

It measures the goodness of fit of the sample to the model class $\mathcal{M}_k$. This term decreases when the complexity of the model increases. The second term of the information criterion, called the *penalty term*, is the number of free parameters of the model. This penalizes too complex models: it increases with the model complexity. Thus, the selected model has a good tradeoff between good description of the data and the model complexity.

For AR models, AIC is asymptotically (i.e., as the sample size tends to infinity) identical to the FPE criterion (Akaike 1972, 1974). Therefore, the AIC estimator also overestimates the true structure parameter asymptotically with positive probability. Shibata (1976) derived the exact asymptotic distribution of the order selected by these estimators.

The classical *cross-validation* principle can be adopted to the model selection problem (e.g., Stone, 1974). The general principle requires dividing the sample set into two subsets, and performing the model estimation based on one subset only. Using the other subset, the candidate model can be validated correctly, that is, the estimation and the validation will be independent. A formulation of this principle for the polynomial fitting problem is the following. Divide the n'th sample $x(n) = \{\, x_t,\, t = t_1, \ldots, t_n \,\}$ into subsets via leaving out the p'th element: $x(n) = x_{\setminus p} \cup \{x_{t_p}\}$, where $x_{\setminus p} = x(n) \setminus \{x_{t_p}\}$. Estimate the coefficients of the polynomial of degree $k-1$ based on the sample set $x_{\setminus p}$:

$$\hat{\theta}_k^{(p)} = \arg\min_{\theta_k \in \Theta_k} \sum_{i \in \{1,\ldots,n\} \setminus \{p\}} \left(x_{t_i} - \left(\theta_k[0] + \theta_k[1]\, t_i + \theta_k[2]\, t_i^2 + \cdots + \theta_k[k-1]\, t_i^{k-1} \right) \right)^2 ,$$

and validate it based on the p'th sample element x_{t_p}:

$$e_k(p) = x_{t_p} - \left(\theta_k[0] + \theta_k[1]\, t_p + \theta_k[2]\, t_p^2 + \cdots + \theta_k[k-1]\, t_p^{k-1} \right) .$$

Calculate this prediction error for all p, and minimize

$$e_k^2 = \sum_{p=1}^{n} e_k(p)^2$$

over the hypothetical k's to obtain the estimated degree of the polynomial. Stone (1977) showed that the cross-validation criterion is asymptotically equivalent to the AIC.

1.3 Consistent model selection

In this work the goodness of model selection will be considered only from the asymptotical point of view; in the literature, this aspect is in the focus.

An estimator $\hat{k}(x(n))$ of the structure parameter k based on the sample $x(n)$ is said to be *consistent* if the probability that the estimator equals the true structure parameter k_0 approaches 1 when the sample size n tends to infinity:

$$Q\left(\hat{k}(x(n)) = k_0\right) \longrightarrow 1 \qquad \text{if } n \to \infty.$$

The estimator $\hat{k}(x(n))$ is said to be *strongly consistent* if it equals the true structure parameter k_0 eventually almost surely as the sample size n tends to infinity:

$$\hat{k}(x(n)) = k_0, \quad \text{eventually almost surely as } n \to \infty.$$

Here and in the sequel, "eventually almost surely" means that with probability 1 there exists a threshold n_0 (depending on the realization $\{\, x_t,\, t \in T \,\}$) such that the claim holds for all $n \geq n_0$.

For the case of density estimation, when the feasible density functions belong to *exponential families*, Schwarz (1978) derived an information criterion from the asymptotic approximation of the Bayesian Maximum A-posteriori Probability (MAP) estimator. Suppose the model class $\mathcal{M}_k$ consists of density functions

$$f_{\theta_k}(x_i) = \exp\left(\, \langle\, \theta_k, y_k(x_i) \,\rangle - b_k(\theta_k) \,\right), \quad \theta_k \in \Theta_k,$$

where $\langle\, \cdot\,,\, \cdot \,\rangle$ denotes the inner product in the $d_k = \dim \Theta_k$ dimensional Euclidean space, $y_k : \mathbb{R} \to \mathbb{R}^{d_k}$ are given functions, and

$$b_k(\theta_k) = \log \int \exp\left(\, \langle\, \theta_k, y_k(x_i) \,\rangle \,\right) \mathrm{d}x_i.$$

Here k ranges over a finite set $\mathcal{K}$. A prior distribution of the parameter vector can be written in the form $\mu = \sum_{k \in \mathcal{K}} \alpha_k \mu_k$, where α_k is the a priori probability that a model with structure parameter k is the true one, and μ_k is the conditional a priori distribution of θ_k under the condition that the true structure parameter is k; μ_k is concentrated on Θ_k. Schwarz showed that, under regularity conditions, the MAP estimator of the parameter vector θ_k from an i.i.d. sample x_1^n asymptotically does not depend on μ, and is equivalent to the maximum likelihood estimator $\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} f_{\theta_k}^{(n)}(x(n))$ in the model class $\mathcal{M}_k$ whose structure parameter k minimizes the value

$$\mathrm{BIC}_k(x(n)) = -\log f_{\hat{\theta}_k}^{(n)}(x(n)) + \frac{\dim \Theta_k}{2} \log n$$

over the set $\mathcal{K}$. This value is called *Bayesian Information Criterion (BIC)*.

The consistency of the BIC estimator in the above situation has been proved by Haughton (1988). Note that for the polynomial fitting problem, Akaike (1977)

introduced the same information criterion with the same notation BIC, in a heuristic way.

For the AR model of order k the Bayesian Information Criterion has the following form:

$$\mathrm{BIC}_k(x_1^n) = -\log f_{\hat{\theta}_k}^{(n)}(x_1^n) + \frac{k}{2}\log n.$$

Hannan and Quinn (1979) proved that the BIC estimator of the order of AR processes is strongly consistent. For the ARMA model of order (p, q) the BIC has the similar form, but k is replaced by $p + q$. Hannan (1980) proved that also in this case, the BIC estimator is strongly consistent.

For the Markov chain of order k, we have

$$\mathrm{BIC}_k(x_1^n) = -\log P_{\hat{\theta}_k}^{(n)}(x_1^n) + \frac{(|A|-1)|A|^k}{2}\log n.$$

Finesso (1992) proved that this gives a strongly consistent order estimator.

It should be emphasized that all consistency results above include the assumption that the number of candidate model classes is finite. This means that there is a known upper bound K on the true order k_0 or (p_0, q_0), and the minimization of the value BIC is for the candidate orders $k \leq K$ or $p \leq K[1]$, $q \leq K[2]$.

1.4 Information theoretical approach

Rissanen (1978, 1983a, 1989) suggested an information theoretical approach to the model selection problem. According to the *Minimum Description Length (MDL)* principle, the best model of the process based on the observed data is the one that gives the shortest description of the observed data, taking into account that the model itself must also be described.

Let each model class $\mathcal{M}_k$ be assigned a uniquely decodable, variable-length binary *code* $C_k^{(n)} : x(n) \mapsto b(x(n))$ which maps a sample $x(n)$ to a binary sequence b whose length can vary with $x(n)$. The codelength function $L_k^{(n)}(x(n))$ is the length of the binary sequence $C_k^{(n)}(x(n))$. Moreover, let $C : k \mapsto b(k)$ be a code of the model classes $\mathcal{M}_k$ which maps a structure parameter k to a binary sequence b. Its codelength function will be denoted by $L(k)$. Thus, using a model class $\mathcal{M}_k$, the sample $x(n)$ can be encoded by encoding $x(n)$ with $C_k^{(n)}(x(n))$ and adding this a preamble $C(k)$ to identify $\mathcal{M}_k$. The MDL criterion is the total length of this description:

$$\mathrm{MDL}_k(x(n)) = L_k^{(n)}(x(n)) + L(k).$$

The MDL estimator selects the model class which provides the shortest description of the sample:

$$\hat{k}(x_1^n) = \arg\min_{k\in\mathcal{K}} \mathrm{MDL}_k(x(n)).$$

Assume for simplicity that A is finite and also that each model M_{θ_k} uniquely determines a hypothetical distribution of the process; as before, its marginal for the sample $x(n)$ is denoted by $P_{\theta_k}^{(n)}$. A uniquely decodable, variable-length binary code $C_k^{(n)}$ can be simply represented by a *coding distribution* $P_k^{(n)}$. To see this, note the well-known fact that $L_k^{(n)}$ is the codelength function of some uniquely decodable code $C_k^{(n)}$ if and only if it satisfies the Kraft inequality $\sum_{x(n)} 2^{-L_k^{(n)}(x(n))} \leq 1$. We may assume that here the equality holds, for otherwise the code could be improved by shortening some codewords. Clearly, for any code $C_k^{(n)}$ with codelength $L_k^{(n)}$ which satisfies the Kraft inequality with the equality, we can write $P_k^{(n)}(x(n)) = 2^{-L_k^{(n)}(x(n))}$. On the other hand, for any probability distribution $P_k^{(n)}$ we can construct a uniquely decodable code $C_k^{(n)}$ with codelength $L_k^{(n)}(x(n)) = \left\lceil -\log_2 P_k^{(n)}(x(n)) \right\rceil$, called a Shannon code. The code determined by the coding distribution $P_k^{(n)}$ will be referred to as $P_k^{(n)}$-code. It should be chosen to be optimal in some sense under the assumption that the true model M_{θ_0} is in the model class $\mathcal{M}_k$. Note, however, that $P_k^{(n)}$ will typically differ from each $P_{\theta_k}^{(n)}$ in the model class $\mathcal{M}_k$.

The *redundancy* of a $P_k^{(n)}$-code relative to the true distribution $P_{\theta_0}^{(n)}$ is

$$R_{\theta_0}^{(n)}(x(n)) = -\log P_k^{(n)}(x(n)) + \log P_{\theta_0}^{(n)}(x(n)).$$

This is the difference of the codelength due to using the coding distribution $P_k^{(n)}$ instead of the true $P_{\theta_0}^{(n)}$. Since the redundancy is a function of the sample, to evaluate the goodness of $P_k^{(n)}$ one usually considers either the maximum of $R_{\theta_0}^{(n)}(x(n))$ for all possible $x(n)$, or its expectation with respect to $P_{\theta_0}^{(n)}$. Moreover, since the true distribution $P_{\theta_0}^{(n)}$ is unknown, as an optimality criterion for $P_k^{(n)}$ under the assumption $M_{\theta_0} \in \mathcal{M}_k$ it is usual to consider worst case maximum or expected redundancy for all feasible distributions $P_{\theta_k}^{(n)}$, $\theta_k \in \Theta_k$, in the role of $P_{\theta_0}^{(n)}$.

For the model class $\mathcal{M}_k$, the worst case maximum redundancy of a $P_k^{(n)}$-code is

$$R^{(n)*} = \sup_{\theta_k\in\Theta_k} \max_{x(n)} R_{\theta_k}^{(n)}(x(n)).$$

It is easy to show that the coding distribution $P_k^{(n)}$ minimizing this quantity is

the *Normalized Maximum Likelihood (NML)* distribution defined as

$$\mathrm{NML}_k^{(n)}(x(n)) = P_{\hat{\theta}_k}^{(n)}(x(n)) \Bigg/ \sum_{x'(n)} P_{\hat{\theta}_k}^{(n)}(x'(n)) \,,$$

where $\hat{\theta}_k = \arg\max_{\theta_k \in \Theta_k} P_{\theta_k}^{(n)}(x(n))$ is the maximum likelihood estimator of the parameter vector θ_k in the model class $\mathcal{M}_k$. Using this coding distribution we get the MDL criterion

$$\mathrm{MDL}_k(x(n)) = -\log P_{\hat{\theta}_k}^{(n)}(x(n)) + \log \left(\sum_{x'(n)} P_{\hat{\theta}_k}^{(n)}(x'(n)) \right) + L(k).$$

Shtarkov (1977) showed that for the case of Markov chains the middle term is asymptotically (as $n \to \infty$, with k fixed) equal to $(1/2)\,(\dim \Theta_k) \log n$. The same holds also in other cases, under suitable regularity conditions, see Rissanen (1996). Hence, when the number of the candidate model classes is finite, the NML version of the MDL criterion is equivalent to BIC.

For the model class $\mathcal{M}_k$, the worst case expected redundancy of a $P_k^{(n)}$-code is

$$\begin{aligned} \bar{R}^{(n)} &= \sup_{\theta_k \in \Theta_k} \mathbb{E}_{\theta_k} \left\{ R_{\theta_k}^{(n)}(x(n)) \right\} = \sup_{\theta_k \in \Theta_k} \sum_{x(n)} P_{\theta_k}^{(n)}(x(n)) \, \log \frac{P_{\theta_k}^{(n)}(x(n))}{P_k^{(n)}(x(n))} \\ &= \sup_{\theta_k \in \Theta_k} \mathrm{D}\left(P_{\theta_k}^{(n)} \middle\| P_k^{(n)} \right), \end{aligned}$$

where $\mathbb{E}_{\theta_k}\{\,\cdot\,\}$ denotes the expected value with respect to the distribution $P_{\theta_k}^{(n)}$, and $\mathrm{D}(\,\cdot\,\|\,\cdot\,)$ is the Kullback – Leibler information divergence.

Concentrating on the Markov chain case, consider first the model class $\mathcal{M}$ equal to $\mathcal{M}_k$ with $k = 0$, the class of i.i.d. processes. In this case, while the exact minimizer of the worst case expected redundancy is unknown, a good coding distribution $P^{(n)}$ is the *Krichevsky – Trofimov (KT)* distribution (Krichevsky and Trofimov, 1981). Its worst case expected redundancy over $\mathcal{M}$ approaches the minimum up to a constant not depending on n. The KT distribution is defined as the mixture of all i.i.d. distributions $P_\theta^{(n)}$, with respect to the Dirichlet distribution ν of parameters 1/2:

$$\mathrm{KT}_0(x_1^n) = \int P_\theta^{(n)}(x_1^n) \; \nu(\mathrm{d}\theta).$$

Direct calculation gives the following explicit expression:

$$\mathrm{KT}_0(x_1^n) = \frac{\prod_{a:N_n(a)\geq 1} \left[\left(N_n(a) - \frac{1}{2}\right) \left(N_n(a) - \frac{3}{2}\right) \cdots \left(\frac{1}{2}\right) \right]}{\left(n - 1 + \frac{|A|}{2}\right) \left(n - 2 + \frac{|A|}{2}\right) \cdots \left(\frac{|A|}{2}\right)},$$

where $N_n(a)$ denotes the number of occurrences of $a \in A$ in the sample x_1^n.

For the Markov chains of order k we have a similar result. For the model class $\mathcal{M}_k$ a good coding distribution is the Krichevsky – Trofimov distribution of order k, denoted by KT_k. It is a mixture of all distributions of form

$$P_{\theta_k}^{(n)}(x_1^n) = \frac{1}{|A|^k} \prod_{i=k+1}^{n} P(x_i \,|x_{i-k}^{i-1}),$$

see (1.1) with $Q(X_1^k = x_1^k) = |A|^{-k}$, where the parameter vector θ_k specifies the matrix of transition probabilities $P(a|\, a_1^k)$, $a \in A$, $a_1^k \in A^k$. The mixing distribution is defined by letting the rows $\left\{ P_{\theta_k}(a|\, a_1^k),\, a \in A \right\}$ of this matrix independent and having the Dirichlet distribution ν as above. We also have an explicit form:

$$\mathrm{KT}_k(x_1^n) = \frac{1}{|A|^k} \prod_{a_1^k : N_n(a_1^k) \geq 1} \frac{\prod_{a_{k+1} : N_n(a_1^{k+1}) \geq 1} \left[\left(N_n(a_1^{k+1}) - \frac{1}{2}\right) \left(N_n(a_1^{k+1}) - \frac{3}{2}\right) \cdots \left(\frac{1}{2}\right) \right]}{\left(N_n(a_1^k) - 1 + \frac{|A|}{2}\right) \left(N_n(a_1^k) - 2 + \frac{|A|}{2}\right) \cdots \left(\frac{|A|}{2}\right)},$$

where $N_n(a_1^k)$ denotes the number of occurrences of $a_1^k \in A^k$ in the sample x_1^n.

The KT_k distribution can be calculated recursively in the sample size n as

$$\mathrm{KT}_k(x_1^n) = \frac{N_{n-1}(x_{n-k}^n) + 1/2}{N_{n-1}(x_{n-k}^{n-1}) + |A|/2}\ \mathrm{KT}_k(x_1^{n-1}).$$

The NML and KT versions of the MDL criterion are asymptotically equivalent, because for the minimzers of the worst case maximum and expected redundancy we have

(1.2)
$$\frac{(|A|-1)|A|^k}{2} \log n - K_1 \leq \min_{P_k^{(n)}} \bar{R}^{(n)} \leq \min_{P_k^{(n)}} R^{(n)*} \leq \frac{(|A|-1)|A|^k}{2} \log n - K_2,$$

where K_1 and K_2 are constants (depending on k).

The MDL estimator can be regarded as a Bayesian MAP estimator when, as above, the coding distribution $P_k^{(n)}$ is a mixture of the distributions $P_{\theta_k}^{(n)}$, $\theta_k \in \Theta_k$, with a suitable mixing distribution $\mu_k^{(n)}$ defined on Θ_k, that is,

$$P_k^{(n)}(x(n)) = \int_{\Theta_k} P_{\theta_k}^{(n)}(x(n))\ \mu_k^{(n)}(\mathrm{d}\theta_k).$$

Indeed, representing the code $C(k)$ of the model class $\mathcal{M}_k$ with the coding distribution $P(k) = 2^{-L(k)}$, minimization of the description length

$$L_k^{(n)}(x(n)) + L(k) = -\log P_k^{(n)}(x(n)) - \log P(k)$$

is equivalent to maximization of $P(k)\,P_k^{(n)}(x(n))$. The latter quantity is proportional to the posterior probability of the structure parameter k, that is, of the conditional probability of k given the sample $x(n)$.

The MDL principle can be extended to the case of general A, say $A = \mathbb{R}$, via discretization, and this leads to similar results as above, see Rissanen (1989), in particular, for AR processes Hemerly and Davis (1989), and for ARMA processes Gerencsér (1987).

1.5 Motivation of the new results

For various processes it has been proved that BIC and MDL estimators of the structure parameter are strongly consistent. This means that the minimizer of BIC or MDL criterion over the candidate structure parameters is equal to the true structure parameter, eventually almost surely as the sample size tends to infinity. Most consistency proofs in the literature include the assumption that the number of candidate structure parameters is finite, that is, there is a known *prior bound* on the structure parameter. This assumption is of technical nature and it simplifies the proof. However, it is undesirable, because in practice usually there is no prior information on the structure parameter, moreover when we have increasing amount of data, we would require to take into account more and more complex hypothetical model classes as candidate ones. Therefore, it is a reasonable aim to drop the assumption of prior bound on the structure parameter.

Csiszár and Shields (2000) proved that the BIC estimator of the order of Markov chains is strongly consistent even if the assumption of the prior constant bound on the order is dropped and based on the n'th sample x_1^n all possible orders $0 \le k < n$ are considered as candidate orders.

At the same time, Csiszár and Shields (2000) pointed out that the same result cannot hold for the MDL estimator. Consider the i.i.d. process with uniform distribution. This process is a Markov chain of order 0. For the MDL criterion

$$\mathrm{MDL}_k(x_1^n) = -\log P_k^{(n)}(x_1^n) + L(k),$$

where the coding distribution $P_k^{(n)}$ is either $\mathrm{NML}_k^{(n)}$ or KT_k, and the codelength $L(k)$ of the order k satisfies $L(k) = o(k)$, we have

$$\hat{k}(x_1^n) = \arg \min_{0 \le k \le \alpha \log n} \mathrm{MDL}_k(x(n)) \ \to +\infty \quad \text{as } n \to \infty,$$

where $\alpha = 4/\log|A|$. This counterexample shows that the MDL estimator fails to be consistent when the prior bound on the order is totally dropped.

Csiszár (2002) proved strong consistency of the MDL estimator of the order of Markov chains when the set of candidate orders is allowed to extend as the sample size n increases, namely, the bound on the orders taken into account is $o(\log n)$ in the KT case, and $\alpha \log n$ with $\alpha < 1/\log|A|$ in the NML case. Let us observe that these MDL estimators need no prior bound on the true order. The consistency was proved for the MDL criterion without the term $L(k)$, which is a stronger result.

The dissertation addresses the model selection problem for two models described below. Strongly consistent estimators of the structure parameters will be presented. Motivated by the above results, the number of candidate model classes will be allowed to grow with the sample size, thus no prior bound on the structure parameter is required.

1.6 The first new result

The model called *tree source* or variable length Markov chain is a refinement of the Markov chain model. Given a Markov chain of order k, for a sequence $a_1^k \in A^k$ the transition probabilities $Q(a|\, a_1^k)$, $a \in A$ may depend not on the whole sequence $a_1, \ldots, a_k$, but only on a subsequence $a_l, \ldots, a_k$; this admits a more parsimonious parameterization.

Consider a process with $T = \mathbb{Z}$ and $|A| < \infty$. For simplicity, assume that all finite dimensional marginals of the distribution Q of the process is strictly positive. The string $s = a_{-l}^{-1} \in A^l$ is a *context* for the process Q if

$$Q(X_i = a \mid X_{-\infty}^{i-1} = x_{-\infty}^{i-1}) = Q(a|\, s) \quad \text{for all } i \in \mathbb{Z},\ a \in A,$$

with suitable transition probabilities $Q(\,\cdot\,|\, s)$, whenever $x_{i-l}^{i-1} = a_{-l}^{-1}$, and no substring $a_{-l'}^{-1}$, $l' < l$ has this property. The set of all contexts is called *context tree*, it will be denoted by $\mathcal{T}$. Assume that for every past sequence $x_{-\infty}^{i-1}$ there exists a context s of finite length l, and the supremum of these lengths is a finite number k. A process with such context tree $\mathcal{T}$ is a Markov chain of order k, but a collection of $(|A|-1)|\mathcal{T}|$ transition probabilities suffices to describe it, instead of $(|A|-1)|A|^k$ ones required for a general Markov chain of order k.

The term context tree refers to its visualization. The contexts s, written backwards, can be regarded as leaves of a tree, where the path from the root to a leaf is determined by the string s. This context tree is complete, that is, each node except the leaves has exactly $|A|$ children.

For a process with context tree $\mathcal{T}$, the probability of a realization x_1^n can be written as

$$Q(X_1^n = x_1^n) = Q(X_1^k = x_1^k) \prod_{s \in \mathcal{T},\, a \in A} Q(a|\, s)^{N_n(s,a)},$$

where $N_n(s,a)$ denotes the number of that occurrences of $a \in A$ in the sequence x_{k+1}^n when the context $s \in \mathcal{T}$ precedes a. Given the sample x_1^n, the maximum of the second factor above is attained for $Q(a|\, s) = \frac{N_n(s,a)}{N_n(s)}$, $a \in A$, $s \in \mathcal{T}$, where $N_n(s) = \sum_{a \in A} N_n(s,a)$. That is, the maximum likelihood estimates of the transition probabilities are the empirical probabilities, as for Markov chains.

For the class of tree models as above, the context tree $\mathcal{T}$ plays the role of the structure parameter. To the analogy of Markov chains, the Bayesian information criterion for the model class determined by $\mathcal{T}$ is

$$\mathrm{BIC}_{\mathcal{T}}(x_1^n) = -\sum_{s \in \mathcal{T},\, a \in A} N_n(s,a) \log \frac{N_n(s,a)}{N_n(s)} + \frac{(|A|-1)|\mathcal{T}|}{2} \log n.$$

The MDL principle can also be formulated for tree models. One can define as before the NML and KT distributions for the class of tree models with context tree $\mathcal{T}$, here we concentrate on the KT distribution. This has the following explicit form:

(1.3)

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) = \frac{1}{|A|^k} \prod_{s: N_n(s) \geq 1} \frac{\prod_{a: N_n(s,a) \geq 1} \left[\left(N_n(s,a) - \frac{1}{2}\right) \left(N_n(s,a) - \frac{3}{2}\right) \cdots \left(\frac{1}{2}\right) \right]}{\left(N_n(s) - 1 + \frac{|A|}{2}\right) \left(N_n(s) - 2 + \frac{|A|}{2}\right) \cdots \left(\frac{|A|}{2}\right)}.$$

We can calculate it recursively in n as

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) = \frac{N_{n-1}(s, x_n) + 1/2}{N_{n-1}(s) + |A|/2}\, \mathrm{KT}_{\mathcal{T}}(x_1^{n-1}),$$

where s is the context for the past $x_{-\infty}^{n-1}$. Similarly as for the class of k'th order Markov chains, the coding distribution $\mathrm{KT}_{\mathcal{T}}$ minimizes the worst case expected redundancy for the class of tree models with context tree $\mathcal{T}$, up to an additive constant. Moreover, the bounds (1.2) also hold if $|A|^k$ is replaced by $|\mathcal{T}|$. Note also that there is a simply code C (see, e.g., Willems, Shtarkov and Tjalkens, 1995) which describes a context tree $\mathcal{T}$ with codelength

$$L(\mathcal{T}) = \frac{|A||\mathcal{T}| - 1}{|A| - 1}. \tag{1.4}$$

For model selection in the tree model setting, instead of using the information criteria above, mostly variants of Rissanen's "context" algorithm (1983b) have been used. The reason is computational complexity: as the number of possible

context trees of depth at most D is very large if D is large, it is not feasible to calculate an information criterion for all candidate context trees and choose the context tree with minimal value. This problem has been partially overcome by Willems, Shtarkov and Tjalkens (1993, 2000). They proved that the MDL estimator with the KT coding distribution (1.3) and the context tree codelength (1.4) is strongly consistent, when there is a prior bound D on the depth of candidate context trees. At the same time they presented an algorithm, called *Context Tree Maximizing (CTM)* method, to calculate the estimator without actually computing and comparing the KT values for all candidate context trees. For the n'th sample x_1^n, the number of elementary computations required by the algorithm is proportional to nD, that is, the algorithm is of linear time.

The first new result of this dissertation, included by Chapter 2, is concerned with the estimation of the context tree, with the definition of the latter significantly generalized. We no longer assume strictly positive probability of all finite sequences, restricting attention in the definition of contexts to those infinite pasts $x_{-\infty}^{i-1}$ for which $Q(X_{i-l}^{i-1} = x_{i-l}^{i-1}) > 0$ for each $l > 0$; thus the context tree may not be complete. More importantly, we allow the case that for some $x_{-\infty}^{i-1}$ as above no context s of finite depth exists, and then the whole past is considered as an infinite context. When the context tree has infinite depth, the process is no longer a Markov chain.

The strong consistency of the BIC estimator and of the KT version of MDL estimator of context trees generalized as above will be proved, defining these estimators with the family of hypothetical context trees allowed to extend as the sample size increases; namely, the depth of context trees considered for sample size n is at most $o(\log n)$. When the true context tree is of infinite depth, it cannot be exactly reconstructed from a finite sample. In this case, we mean by strong consistency that for each $D > 0$, the estimated context tree pruned at level D equals the true one pruned at the same level D, eventually almost surely. In addition, algorithms are presented to calculate both estimators in linear time. The computation complexity of the on-line versions of these algorithms, that is, when they are being computed at all instances while the sample size is increasing, is also discussed.

1.7 The second new result

The second new result of this dissertation addresses the model selection problem for *Markov random fields*. The Gibbs fields are spatial processes which may have

higher dimensional index set. The Markov random fields are Gibbs fields having the Markov property in space.

Let $T = \mathbb{Z}^d$. The points of the d dimensional integer *lattice* $\mathbb{Z}^d$ are called *sites*. Let A be a Polish space (i.e., a complete separable metric space), and let $\mathcal{X} = A^{\mathbb{Z}^d}$. The elements x of the set $\mathcal{X}$ are called configurations of the field. The configuration in the region $V \subseteq \mathbb{Z}^d$ or at the site $i \in \mathbb{Z}^d$ will be denoted by $x(V) \in A^V$, respectively $x(i) \in A$. Let ρ be a probability measure on A, and let $\rho^{\otimes \mathbb{Z}^d}$ be the corresponding product measure on $\mathcal{X}$.

An *interaction potential* I is a family $I = \{\, I_V,\, V \subset \mathbb{Z}^d \text{ finite} \,\}$ of bounded continuous functions $I_V : \mathcal{X} \to \mathbb{R}$ which depend on the configuration in the region V only. Moreover, suppose translation invariance of I:

$$I_V(\tau^j(x)) = I_{V^j}(x) \quad \text{for all } V,\, x \in \mathcal{X},\, j \in \mathbb{Z}^d,$$

where $\tau^j(x) : \mathcal{X} \to \mathcal{X}$ with $(\tau^j(x))(i) = x(i+j)$, and $V^j = \{\, i + j : i \in V \,\}$. Suppose also summability of I:

$$\sum_{V:\, 0 \in V} \left(\sup_{x \in \mathcal{X}} I_V(x) \right) < +\infty,$$

where 0 is the origin in $\mathbb{Z}^d$. Denote $\Lambda \subset \mathbb{Z}^d$ a finite region, and Λ^c its complement. For the interaction potential I, the *energy* of a configuration $x(\Lambda) \in A^\Lambda$, given the boundary condition $y(\Lambda^c) \in A^{\Lambda^c}$, is

$$U_\Lambda(x(\Lambda) \mid y(\Lambda^c)) = \sum_{V : V \cap \Lambda \neq \varnothing} I_V(x(\Lambda) \vee y(\Lambda^c)),$$

where $x(\Lambda) \vee y(\Lambda^c) \in \mathcal{X}$ is the configuration equal to $x(\Lambda)$ on Λ and to $y(\Lambda^c)$ on Λ^c.

Consider the parameter estimation problem, when there is only one model class with parameter vector $\theta \in \Theta \subset \mathbb{R}^k$. In this case we have a known collection of interaction potentials $I^{(l)}$, $l = 1, \ldots, k$ and of energy functions $U^{(l)}$, $l = 1, \ldots, k$. The *Gibbs specification* parameterized by the parameter vector θ is

$$\pi_{\Lambda,\theta}(x(\Lambda) \mid y(\Lambda^c)) = Z_{\Lambda, y(\Lambda^c)}(\theta)^{-1} \exp\left(\sum_{l=1}^{k} \theta[l]\, U_\Lambda^{(l)}(x(\Lambda) \mid y(\Lambda^c)) \right), \tag{1.5}$$

where $Z_{\Lambda, y(\Lambda^c)}(\theta)$ is the normalizing factor

$$Z_{\Lambda, y(\Lambda^c)}(\theta) = \int \exp\left(\sum_{l=1}^{k} \theta[l]\, U_\Lambda^{(l)}(x(\Lambda) \mid y(\Lambda^c)) \right) \rho^{\otimes \Lambda}(\mathrm{d}x(\Lambda)). \tag{1.6}$$

The set $\mathcal{G}(\theta)$ of *Gibbs distributions* is the set of all probability distributions P_θ on $\mathcal{X}$ whose conditional finite marginals, given the boundary condition y, satisfy

$$P_\theta(\mathrm{d}x(\Lambda) \mid y(\Lambda^c)) = \pi_{\Lambda,\theta}(x(\Lambda) \mid y(\Lambda^c))\ \rho^{\otimes\Lambda}(\mathrm{d}x(\Lambda))$$

for all finite regions $\Lambda \subset \mathbb{Z}^d$ and P_θ-almost every $y(\Lambda^c)) \in A^{\Lambda^c}$. The translation invariance and summability assumptions imply that there exists such distribution P_θ. On the other hand, the above equation, called Dobrushin – Lanford – Ruelle equation, is not necessarily satisfied by only one distribution P_θ, due to the fact that a distribution is not necessarily determined by its conditional marginals. This phenomenon is called *phase transition*.

Markov random fields are Gibbs distributions with finite range interaction potentials, which means that there exists a positive constant such that $I_V \equiv 0$ for all V whose diameter is greater than this constant. In this case the conditional probability density $P_\theta(\mathrm{d}x(\Lambda)|\, y(\Lambda^c))$ depends on the value of the boundary condition y at a finite number of sites only.

The parameter estimation problem consists in estimating the parameter vector $\theta \in \Theta$ based on a realization $x \in \mathcal{X}$ observed in a finite region Λ_n; thus the n'th sample is $x(\Lambda_n)$. For simplicity, it is usually assumed that $\Lambda_n = [-n,\, n]^d$. Similarly to the case of Markov chains, likelihood is defined via the conditional distribution of the sample given some boundary condition. The maximum likelihood estimator is the parameter vector minimizing the likelihood function $\pi_{\Lambda_n,\theta}(x(\Lambda_n)|\, y(\Lambda_n^c))$ for some boundary condition y, over the set Θ. An alternative of fixing the boundary condition y is shrinking the region Λ_n to $\bar{\Lambda}_n \subset \Lambda_n$ such that $x(\Lambda_n \backslash \bar{\Lambda}_n)$ determines $\pi_{\bar{\Lambda}_n,\theta}(x(\bar{\Lambda}_n)|\, y(\bar{\Lambda}_n^c))$. Here the former option is chosen, while the second new result of dissertation adopts the latter.

The maximum likelihood estimator can not be calculated in practice, because of the intractable integrals in (1.6). The *pseudo-likelihood*, introduced by Besag (1974), is defined as

$$\mathrm{PL}_\theta(x(\Lambda_n) \mid y(\Lambda_n^c)) = \prod_{i\in\Lambda_n} \pi_{\{i\},\theta}\left(x(i) \mid (x(\Lambda_n) \vee y(\Lambda_n^c))\left(\{i\}^c\right)\right). \tag{1.7}$$

The maximum pseudo-likelihood estimator of the parameter vector $\theta \in \Theta$ based on the sample $x(\Lambda_n)$, given the boundary condition y, is

$$\hat{\theta}(x(\Lambda_n) \mid y(\Lambda_n^c)) = \arg\max_{\theta\in\Theta} \mathrm{PL}_\theta(x(\Lambda_n) \mid y(\Lambda_n^c)).$$

This estimator can be calculated favorably, because the form (1.6) contains only a single integral for $\Lambda = \{i\}$.

Comets (1992) proved that when the model is identifiable, that is, $\mathcal{G}(\theta) \cap \mathcal{G}(\theta_0) = \varnothing$ for all θ, θ_0 with $\theta \neq \theta_0$, the maximum pseudo-likelihood estimator is strongly consistent, which means that

$$\hat{\theta}(x(\Lambda_n) \mid y(\Lambda_n^c)) \longrightarrow \theta_0,$$

P_θ-almost surely as $n \to \infty$, for any boundary condition y.

The second new result of dissertation, included by Chapter 3, is concerned with model selection problem for Markov random fields when A is finite. In this case the Markov random field can be parameterized differently from the above, namely, by the specification (1.5) for the region consisting of one site. That is, the parameters describe the (translation invariant) conditional distribution at a site, say at the origin $0 \in \mathbb{Z}^d$, given the values $x(j)$ at all of the other sites $j \in \mathbb{Z}^d \backslash \{0\}$. The structure parameter is the smallest region $\Gamma \subset \mathbb{Z}^d \backslash \{0\}$ in which the values $x(\Gamma)$ of the sites determine this conditional distribution. Therefore, for the structure parameter $\Gamma \subset \mathbb{Z}^d$, $0 \notin \Gamma$ the parameter vector is the matrix

$$P_\Gamma = \left\{ P_\Gamma(x(0) \mid x(\Gamma)),\, x(0) \in A,\, x(\Gamma) \in A^\Gamma \right\}.$$

Note that the number of free parameters is not equal to $(|A| - 1)|A|^{|\Gamma|}$, because the conditional probabilities may not be chosen arbitrarily.

The region Γ is called *basic neighborhood*. Its finiteness is equivalent to that the Gibbs distribution is a Markov random field. On account of (1.5), the conditional probability of a configuration in a finite region Λ given the boundary condition $x(\Lambda^c)$ depends on $x(\Lambda_\Gamma)$ only, where $\Lambda_\Gamma = (\cup_{i \in \Lambda} \Gamma^i) \setminus \Lambda$.

With this parameterization the pseudo-likelihood (1.7) can be written as

$$\mathrm{PL}_\Gamma(x(\Lambda_n),\, P_\Gamma) = \prod_{a(\Gamma \cup \{0\}) \in A^{\Gamma \cup \{0\}}} P_\Gamma(a(0) \mid a(\Gamma))^{N_n(a(\Gamma \cup \{0\}))},$$

where $N_n(a(\Gamma \cup \{0\}))$ denotes the number of occurrences of the configuration $a(\Gamma \cup \{0\})$ in the sample $x(\Lambda_n)$. Below we consider the above PL_Γ defined for all matrices P_Γ with non-negative components satisfying $\sum_{a(0) \in A} P_\Gamma(a(0) |\, a(\Gamma)) = 1$ for all $a(\Gamma) \in A^\Gamma$, whether or not P_Γ is the one-point specification of some Gibbs distribution. Then the maximum of pseudo-likelihood is attained for $P_\Gamma(a(0) |\, a(\Gamma)) = \frac{N_n(a(\Gamma \cup \{0\}))}{N_n(a(\Gamma))}$, where $N_n(a(\Gamma)) = \sum_{a \in A} N_n(a(\Gamma \cup \{0\}))$. Thus, the *maximum pseudo-likelihood* is

$$\mathrm{MPL}_\Gamma(x(\Lambda_n)) = \prod_{a(\Gamma \cup \{0\}) \in A^{\Gamma \cup \{0\}}} \left[\frac{N_n(a(\Gamma \cup \{0\}))}{N_n(a(\Gamma))} \right]^{N_n(a(\Gamma \cup \{0\}))}.$$

Our aim is to estimate the true basic neighborhood Γ_0 based on the sample $x(\Lambda_n)$. To this end, we use a modification of the Bayesian information criterion, with maximum likelihood replaced by maximum pseudo-likelihood and the number of free parameters by a quantity proportional to the number of parameters. The so obtained information criterion, called *Pseudo-Bayesian Information Criterion (PIC)*, is

$$\mathrm{PIC}_\Gamma(x(\Lambda_n)) = -\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) + |A|^{|\Gamma|} \log |\Lambda_n|,$$

where $|\Lambda_n|$ is the sample size, the cardinality of the region Λ_n. The PIC estimator $\hat{\Gamma}(x(\Lambda_n))$ is the basic neighborhood minimizing this value over the set of hypothetical Γ's.

It will be proved that the PIC estimator of the basic neighborhood is strongly consistent, that is, $\hat{\Gamma}(x(\Lambda_n)) = \Gamma_0$ eventually almost surely as $n \to \infty$, even when the set of candidate basic neighborhoods is allowed to grow with the sample size $|\Lambda_n|$; namely, the diameter of the basic neighborhoods considered based on the sample $x(\Lambda_n)$ is $o(\log^{1/(2d)} |\Lambda_n|)$, at most. The sample region Λ_n is not required to be the cube $[-n, n]^d$. Moreover, the presence of phase transition does not effect the result.

Chapter 2

Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL

2.1 Introduction

In this chapter, *process* always means a stationary ergodic stochastic process with finite alphabet. Processes are often described by the collection of the conditional probabilities of the possible symbols given the infinite pasts. When these probabilities depend on at most k previous symbols, the process is a Markov chain of order k.

The number of parameters of a general Markov chain grows exponentially with the order. A more efficient description is possible if the strings determining the conditional probabilities – referred to as *contexts* – are of variable length, sometimes substantially shorter than the order k. Models of this kind, and the term *context tree* (referring to the representation of the set of contexts as a tree) dates back to Rissanen (1983a). These models are also called finite memory sources or tree sources (Weinberger, Lempel and Ziv, 1992), (Weinberger, Rissanen and Feder, 1995), (Willems, Shtarkov and Tjalkens, 1995) or variable length Markov chains (Bühlmann and Wyner, 1999). We note that the terms context and context tree appear in the literature in various senses. Here, the context tree of a finite memory process means, in effect, the minimal tree admitting a tree source representation of the process; the exact definition will be given in Section 2.2.

As indicated above, the context tree model is typically used to more efficiently describe certain Markov chains (of finite order k) and, accordingly, the context

tree has finite depth k. Also, the context tree is usually required to be complete, that is, each internal node has as many children as the cardinality of the alphabet.

In this work, we drop the requirements of finite depth and completeness. The term "infinite-depth context tree" appears in (Willems, 1998), but it involves an "indeterminate symbol" ε such that infinitely many ε's may precede a finite number of symbols of the true alphabet. Our approach does not involve such ε, and processes whose context trees have infinite depth are no longer Markov chains. Dropping the completeness requirement is justified by disregarding strings of zero probability. Not necessarily complete context trees were previously considered in (Martín, Seroussi and Weinberger, 2004), in a sense different from ours.

We address the problem of statistical estimation of the context tree in the indicated generality, based on an observed finite realization of the process. This task, for finite depth context trees, has been considered, among others, in the references above. As distinct from those, here we show that the standard methods of Bayesian Information Criterion (BIC) of Schwarz (1978) and the Krichevsky-Trofimov (KT) version of Minimum Description Length (MDL) of Rissanen (1989), (Barron, Rissanen and Yu, 1998), are fully appropriate for this purpose. Note that BIC is commonly regarded as an approximation of MDL, but this is justified only when a fixed finite number of model classes is considered, see (Csiszár and Shields, 2000).

For order estimation of Markov chains, it is well known that BIC and both the KT and Normalized Maximum Likelihood (NML) versions of MDL are strongly consistent when the number of candidate model classes is finite, that is, when there is a known upper bound on the order (Finesso, 1992). The consistency of the BIC order estimator without such prior bound has been proved by Csiszár and Shields (2000). That paper also contains a counterexample to the consistency of the KT and NML estimators without any bound on the order, or with a bound depending on the sample size n, equal to a sufficiently large constant times $\log n$. The consistency of the KT and NML order estimators was proved by Csiszár (2002) with bound $o(\log n)$ resp. $O(\log n)$ on the order.

For the estimation of context trees of finite memory processes, in the literature mostly variants of Rissanen's "context" algorithm (1983b) are used. In particular, Bühlmann and Wyner (1999) proved the consistency of such an algorithm not assuming any prior bound on the depth of the context tree (but using a bound allowed to grow with the sample size). The KT version of MDL has been applied to context tree estimation by Willems, Shtarkov and Tjalkens (1993, 2000), and its consistency was proved assuming a known upper bound on the depth.

In addition to consistency, an important feature of estimators is their computational complexity. One reason for not having used standard statistical methods for context tree estimation was the computational infeasibility of comparing a very large number of hypothetical models. As shown in (Willems, Shtarkov and Tjalkens, 1993, 2000), however, such time-consuming comparisons can be avoided by clever use of tree techniques, viz. the Context Tree Maximizing (CTM) method, and context tree estimation via the KT version of MDL can be implemented in linear time (assuming a known upper bound on the depth). Recent results on context tree estimation in linear time, assuming finite depth but no known upper bound on it, appear in (Baron and Bresler, 2004), (Martín, Seroussi and Weinberger, 2004). We note that much of the literature of context tree models is motivated by universal source coding; in particular, the CTM method is a modification of the celebrated Context Tree Weighting data compression algorithm of Willems, Shtarkov and Tjalkens (1995).

In this chapter we prove that the BIC and KT estimators of the context tree of a process is strongly consistent when the depths of the hypothetical context trees are allowed to grow with the sample size n as $o(\log n)$. Here the context tree may be of infinite depth, and is not necessarily complete. Strong consistency means in the finite depth case that the estimated context tree is equal to the true one, eventually almost surely as $n \to \infty$, while otherwise, that the estimated context tree cut off at any fixed level is equal to the true one cut off at the same level, eventually almost surely as $n \to \infty$. In addition, we provide algorithms to calculate both estimators in $O(n)$ time. These algorithms are based on the CTM method; in particular, BIC also admits a CTM-like implementation.

By our consistency result, if the context tree of a process has finite depth, it can be exactly recovered, with probability 1, when the sample size is large enough. While our result gives no indication how large this sample size should be, a heuristic rule might be to stop when the estimated context tree "stabilizes", that is, it remains unchanged when the sample size n runs over a large interval. The last result in this chapter shows that our context tree estimators can be calculated on-line in such a way that $o(n \log n)$ time suffices to calculate them for all sample sizes $i \leq n$. This implies that the above stopping rule can be implemented with only a small increment in the order of required computations.

The structure of this chapter is the following. In Section 2.2 we introduce the notation and definitions, and formulate the results for the BIC estimator and KT estimator about strong consistency and computational complexity. In Section 2.3 we introduce the algorithms for calculating the estimators, and establish their

claimed computational complexity both for off-line and on-line calculations. In Section 2.4 we prove the consistency theorems. Section 2.5 contains some remarks on the results.

2.2 Notation and statement of the main results

For a finite set A we denote its cardinality by $|A|$. A *string* $s = a_m a_{m+1} \dots a_n$ (with $a_i \in A$, $m \le i \le n$) is denoted also by a_m^n; its length is $l(s) = n - m + 1$. The empty string is denoted by $\varnothing$, its length is $l(\varnothing) = 0$. The concatenation of the strings u and v is denoted by uv. We say that a string v is a *postfix* of a string s, denoted by $s \succeq v$, when there exists a string u such that $s = uv$. For a proper postfix, that is, when $s \neq v$, we write $s \succ v$. A postfix of a semi-infinite sequence $a_{-\infty}^{-1} = \dots a_{-k} \dots a_{-1}$ is defined similarly. Note that in the literature $\succ$ more often denotes the prefix relation.

A set $\mathcal{T}$ of strings, and perhaps also of semi-infinite sequences, is called a *tree* if no $s_1 \in \mathcal{T}$ is a postfix of any other $s_2 \in \mathcal{T}$.

Each string $s = a_1^k \in \mathcal{T}$ is visualized as a path from a leaf to the root (drawn with the root at the top), consisting of k edges labeled by the symbols $a_1 \dots a_k$. A semi-infinite sequence $a_{-\infty}^{-1} \in \mathcal{T}$ is visualized as an infinite path to the root. The strings $s \in \mathcal{T}$ are identified also with the leaves of the tree $\mathcal{T}$, the *leaf* s is the leaf connected with the root by the path visualizing s as above. Similarly, the *nodes* of the tree $\mathcal{T}$ are identified with the finite postfixes of all (finite or infinite) $s \in \mathcal{T}$, the root being identified with the empty string $\varnothing$. The *children* of a node s are those strings as, $a \in A$, that are themselves nodes, that is, postfixes of some $s' \in \mathcal{T}$.

The tree $\mathcal{T}$ is *complete* if each node except the leaves has exactly $|A|$ children. A weaker property we shall need is *irreducibility*, which means that no $s \in \mathcal{T}$ can be replaced by a proper postfix without violating the tree property. The family of irreducible trees will be denoted by $\mathcal{I}$.

We write $\mathcal{T}_2 \succeq \mathcal{T}_1$ for two trees $\mathcal{T}_1$ and $\mathcal{T}_2$, when each $s_2 \in \mathcal{T}_2$ has a postfix $s_1 \in \mathcal{T}_1$, and each $s_1 \in \mathcal{T}_1$ is a postfix of some $s_2 \in \mathcal{T}_2$. When we insist on $\mathcal{T}_2 \neq \mathcal{T}_1$, we write $\mathcal{T}_2 \succ \mathcal{T}_1$.

Denote $d(\mathcal{T})$ the depth of the tree $\mathcal{T}$: $d(\mathcal{T}) = \max\{\, l(s), s \in \mathcal{T} \,\}$. Let $\mathcal{T}\big|_K$ denote the tree $\mathcal{T}$ pruned at level K:

(2.1)

$$\mathcal{T}\big|_K = \{\, s' : s' \in \mathcal{T} \text{ with } l(s') \le K \text{ or } s' \text{ is a } \lfloor K \rfloor\text{-length postfix of some } s \in \mathcal{T} \,\}.$$

Consider a stationary ergodic stochastic process $\{ X_i, -\infty < i < +\infty \}$ with finite alphabet A. Write

$$Q(a_m^n) = \text{Prob}\{ X_m^n = a_m^n \},$$

and, if $s \in A^k$ has $Q(s) > 0$, write

$$Q(a|\, s) = \text{Prob}\{ X_0 = a \mid X_{-k}^{-1} = s \}.$$

A process as above will be referred to as process Q.

Definition 2.1. *A string $s \in A^k$ is a* context *for a process Q if $Q(s) > 0$ and*

$$\text{Prob}\{ X_0 = a \mid X_{-\infty}^{-1} = x_{-\infty}^{-1} \} = Q(a|\, s), \quad \textit{for all } a \in A,$$

whenever s is a postfix of the semi-infinite sequence $x_{-\infty}^{-1}$, and no proper postfix of s has this property. An infinite context *is a semi-infinite sequence $x_{-\infty}^{-1}$ whose postfixes x_{-k}^{-1}, $k = 1, 2, \ldots$ are of positive probability but none of them is a context.*

Clearly, the set of all contexts is a tree. It will be called the *context tree* $\mathcal{T}_0$ of the process Q.

Remark 2.2. The context tree $\mathcal{T}_0$ has to be complete if $Q(s) > 0$ for all strings s. In general, for each node s of the context tree $\mathcal{T}_0$, exactly those as, $a \in A$, are the children of s for which $Q(as) > 0$. Moreover, Definition 2.1 implies that always $\mathcal{T}_0 \in \mathcal{I}$. □

When the context tree has depth $d(\mathcal{T}_0) = k_0 < \infty$, the process Q is a Markov chain of order k_0. In this case the context tree leads to a parsimonious description of the process, because a collection of $(|A|-1)|\mathcal{T}_0|$ transition probabilities suffices to describe the process, instead of $(|A|-1)|A|^{k_0}$ ones. Note that the context tree of an i.i.d. process consists of the root $\varnothing$ only, thus $|\mathcal{T}_0| = 1$.

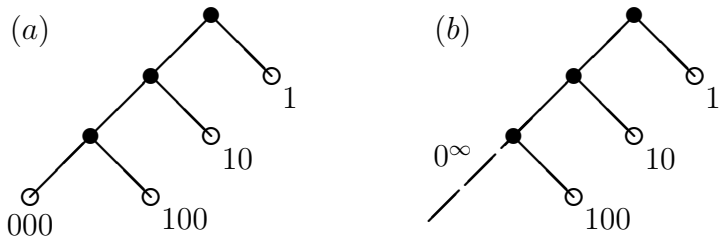

Figure 2.1: Context tree of a renewal process. (*a*) $k_0 = 3$. (*b*) $k_0 = \infty$.

Example 2.3. *(Renewal Process).* Let $A = \{0, 1\}$ and suppose that the distances between the occurrences of 1's are i.i.d. Denote p_j the probability that this distance is j, that is, $p_j = Q(10^{j-1}1)$. Then for $k \geq 1$ we have $Q(10^{k-1}) = \sum_{i=k}^{\infty} p_i \triangleq q_k$, $Q_k = Q(1|\, 10^{k-1}) = p_k/q_k$. Let $Q_0 = Q(1) \triangleq q_0$. Denote k_0 the smallest integer such that Q_k is constant for $k \geq k_0$ with $q_k > 0$, or $k = \infty$ if no such integer exists. Then the contexts are the strings 10^{i-1}, $i \leq k_0$, and the string 0^{k_0} (if $k_0 < \infty$) or the semi-infinite sequence 0^∞ (if $k_0 = \infty$), see Fig. 2.1. □

In this work, we are concerned with the statistical estimation of the context tree $\mathcal{T}_0$ from the sample x_1^n, a realization of X_1^n. We demand strongly consistent estimation. We mean by this in the case $d(\mathcal{T}_0) < \infty$ that the estimated context tree equals $\mathcal{T}_0$ eventually almost surely as $n \to \infty$, while otherwise that the estimated context tree pruned at any fixed level K equals $\mathcal{T}_0\big|_K$ eventually almost surely as $n \to \infty$, see (2.1). Here and in the sequel, "eventually almost surely" means that with probability 1 there exists a threshold n_0 (depending on the doubly infinite realization $x_{-\infty}^{\infty}$) such that the claim holds for all $n \geq n_0$.

Let $N_n(s, a)$ denote the number of occurrences of the string $s \in A^{l(s)}$ followed by the letter $a \in A$ in the sample x_1^n, where s is supposed to be of length at most $\log n$, and – for technical reason – only the letters in positions $i > \log n$ are considered:

$$N_n(s, a) = \left| \left\{ i :\ \log n < i \leq n,\, x_{i-l(s)}^{i-1} = s,\, x_i = a \right\} \right|.$$

Logarithms are to the base e. The number of such occurrences of s is denoted by $N_n(s)$:

$$N_n(s) = \left| \left\{ i :\ \log n < i \leq n,\, x_{i-l(s)}^{i-1} = s \right\} \right|.$$

Given a sample x_1^n, a *feasible tree* is any tree $\mathcal{T}$ of depth $d(\mathcal{T}) \leq \lceil \log n \rceil$ such that $N_n(s) \geq 1$ for all $s \in \mathcal{T}$, and each string s' with $N_n(s') \geq 1$ is either a postfix of some $s \in \mathcal{T}$ or has a postfix $s \in \mathcal{T}$. A feasible tree $\mathcal{T}$ is called *r-frequent* if $N_n(s) \geq r$ for all $s \in \mathcal{T}$. The family of all feasible respectively r-frequent trees is denoted by $\mathcal{F}_1(x_1^n)$ respectively $\mathcal{F}_r(x_1^n)$.

Clearly,

$$\sum_{a \in A} N_n(s, a) = N_n(s), \quad \text{and} \quad \sum_{s \in \mathcal{T}} N_n(s) = n - \lceil \log n \rceil$$

for any feasible tree $\mathcal{T}$. Regarding such a tree $\mathcal{T}$ as a hypothetical context tree, the probability of the sample x_1^n can be written as

$$Q(x_1^n) = Q(x_1^{\lceil \log n \rceil}) \prod_{s \in \mathcal{T},\, a \in A} Q(a|\, s)^{N_n(s,a)}.$$

With some abuse of terminology, for a hypothetical context tree $\mathcal{T} \in \mathcal{F}_1(x_1^n)$ we define the maximum likelihood $\mathrm{ML}_{\mathcal{T}}(x_1^n)$ as the maximum in $Q(a|\, s)$ of the second factor above. Then

$$\log \mathrm{ML}_{\mathcal{T}}(x_1^n) = \sum_{s \in \mathcal{T},\, a \in A} N_n(s,a) \log \frac{N_n(s,a)}{N_n(s)}.$$

We investigate two information criteria to estimate $\mathcal{T}_0$, both motivated by the MDL principle. An information criterion assigns a score to each hypothetical model (here, context tree) based on the sample, and the estimator will be that model whose score is minimal.

Definition 2.4. *Given a sample x_1^n, the* BIC *for a feasible tree $\mathcal{T}$ is*

$$\mathrm{BIC}_{\mathcal{T}}(x_1^n) = -\log \mathrm{ML}_{\mathcal{T}}(x_1^n) + \frac{(|A|-1)|\mathcal{T}|}{2} \log n.$$

Remark 2.5. Characteristic for BIC is the "penalty term" half the number of free parameters times $\log n$. Here, a process Q with context tree $\mathcal{T}$ is described by the conditional probabilities $Q(a|\, s)$, $a \in A$, $s \in \mathcal{T}$, and $(|A|-1)|\mathcal{T}|$ of these are free parameters when the tree $\mathcal{T}$ is complete. On the other hand, for a process with an incomplete context tree, the probabilities of certain strings must be 0, hence the number of free parameters is typically smaller than $(|A|-1)|\mathcal{T}|$ when $\mathcal{T}$ is not complete. Thus, Definition 2.4 involves a slight abuse of terminology. We note that replacing $(|A|-1)/2$ in Definition 2.4 by any $c > 0$ would not affect the results below and their proofs. □

It is known (Csiszár and Shields, 2000) that for estimating the order of Markov chains, the BIC estimator is consistent without any restriction on the hypothetical orders. The Theorem below does need a bound on the depth of the hypothetical context trees. Still, as this bound grows with the sample size n, no a priori bound on the size of the unknown $\mathcal{T}_0$ is required, in fact, even $d(\mathcal{T}_0) = \infty$ is allowed. Note also that the presence of this bound decreases computational complexity.

Theorem 2.6. *In the case $d(\mathcal{T}_0) < \infty$, the BIC estimator*

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \arg \min_{\mathcal{T} \in \mathcal{F}_1(x_1^n) \cap \mathcal{I},\, d(\mathcal{T}) \le D(n)} \mathrm{BIC}_{\mathcal{T}}(x_1^n)$$

with $D(n) = o(\log n)$, satisfies

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \mathcal{T}_0$$

eventually almost surely as $n \to \infty$.

In general case, the BIC estimator

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \arg \min_{\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I},\, d(\mathcal{T}) \le D(n)} \mathrm{BIC}_{\mathcal{T}}(x_1^n)$$

with $D(n) = o(\log n)$ *and arbitrary* $0 < \alpha < 1$*, satisfies for any constant* K

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n)\big|_K = \mathcal{T}_0\big|_K$$

eventually almost surely as $n \to \infty$*.*

Proof. See Section 2.4. □

Remark 2.7. Here and in Theorem 2.9 below, the indicated minimum is certainly attained, as the number of feasible trees is finite, but the minimizer is not necessarily unique; in that case, either minimizer can be taken as arg min. □

The other information criterion we consider is the Krichevsky – Trofimov codelength (see (Krichevsky and Trofimov, 1981), (Willems, Shtarkov and Tjalkens, 1995)). Note that a code with length-function equal to $\mathrm{KT}_{\mathcal{T}}(x_1^n)$ below minimizes the worst case average redundancy, up to an additive constant, for the class of processes with context tree $\mathcal{T}$.

Definition 2.8. *Given a sample* x_1^n*, the* KT *criterion for a feasible tree* $\mathcal{T}$ *is*

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) = -\log P_{\mathrm{KT},\mathcal{T}}(x_1^n),$$

where

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \frac{1}{|A|^{\lceil \log n \rceil}} \prod_{s \in \mathcal{T}} \frac{\prod_{a:N_n(s,a)\ge 1} \left[\left(N_n(s,a) - \frac{1}{2}\right) \left(N_n(s,a) - \frac{3}{2}\right) \cdots \left(\frac{1}{2}\right) \right]}{\left(N_n(s) - 1 + \frac{|A|}{2}\right) \left(N_n(s) - 2 + \frac{|A|}{2}\right) \cdots \left(\frac{|A|}{2}\right)}$$

is the KT-probability of x_1^n *corresponding to* $\mathcal{T}$*.*

For estimating the order of Markov chains, the consistency of the KT estimator has been proved when the hypothetical orders are $o(\log n)$ (Csiszár, 2002), while without any bound on the order, or with a bound equal to a sufficiently large constant times $\log n$, a counterexample to its consistency is known (Csiszár and Shields, 2000).

Theorem 2.9. *In the case* $d(\mathcal{T}_0) < \infty$*, the KT estimator*

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \arg \min_{\mathcal{T} \in \mathcal{F}_1(x_1^n) \cap \mathcal{I},\, d(\mathcal{T}) \le D(n)} \mathrm{KT}_{\mathcal{T}}(x_1^n)$$

with $D(n) = o(\log n)$*, satisfies*

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \mathcal{T}_0$$

eventually almost surely as $n \to \infty$*.*

In general case, the KT estimator

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \arg \min_{\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I},\, d(\mathcal{T}) \le D(n)} \mathrm{KT}_{\mathcal{T}}(x_1^n)$$

with $D(n) = o(\log n)$ *and arbitrary* $1/2 < \alpha < 1$*, satisfies for any constant* K

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n)\big|_K = \mathcal{T}_0\big|_K$$

eventually almost surely as $n \to \infty$*.*

Proof. See Section 2.4. □

Remark 2.10. Strictly speaking, the MDL principle would require to minimize the "codelength" $\mathrm{KT}_{\mathcal{T}}(x_1^n)$ incremented by an additional term, the "codelength of $\mathcal{T}$" (called the cost of $\mathcal{T}$ in (Willems, Shtarkov and Tjalkens, 1995)). This additional term is omitted, since this does not affect the consistency result. □

Corollary 2.11. *The vector of the empirical conditional probabilities,*

$$\widehat{Q}_{\widehat{\mathcal{T}}}(a|\, s) = \frac{N_n(s,a)}{N_n(s)}\,, \quad a \in A,\, s \in \widehat{\mathcal{T}},$$

converges to that of the true conditional probabilities $Q(a|s)$*,* $a \in A$*,* $s \in \mathcal{T}_0$ *almost surely as* $n \to \infty$*, where* $\widehat{\mathcal{T}}$ *is either the BIC estimator or the KT estimator.*

Proof. Immediate from Theorems 2.6, 2.9 and the ergodic theorem. □

In practice, it is unfeasible to calculate estimators via computing the value of an information criterion for each model, since the number of the hypothetical context trees is very large. However, the algorithms in the next section admit finding the considered estimators with practical computational complexity.

As usual, see (Baron and Bresler, 2004), (Martín, Seroussi and Weinberger, 2004), we assume that the computations are done in registers of size $O(\log n)$.

We consider both off-line and on-line methods. Note that on-line calculation of the estimator is useful when the sample size is not fixed but we keep sampling until the estimator becomes "stable", say it remains constant when the sample size is doubled.

Theorem 2.12. *The number of computations needed to determine the BIC estimator and the KT estimator in Theorems 2.6 and 2.9 for a given sample* x_1^n *is* $O(n)$*, and this can be achieved storing* $O(n^\varepsilon)$ *data, where* $\varepsilon > 0$ *is arbitrary.*

Proof. See Section 2.3. □

Theorem 2.13. *Given a sample x_1^n, the number of computations needed to determine the KT estimator in Theorem 2.9 simultaneously for all subsamples x_1^i, $i \le n$, is $o(n \log n)$, and this can be achieved storing $O(n^\varepsilon)$ data at any time, where $\varepsilon > 0$ is arbitrary.*

The same holds for the BIC estimator in Theorem 2.6 with a slightly modified definition of BIC. Namely, let k_m, $m \in \mathbb{N}$ denote the smallest integer k satisfying $D(k) = m$, and replace n in the penalty term in Definition 2.4 by the smallest member of the sequence $\{k_m\}$ larger than n.

Proof. See Section 2.3. □

2.3 Computation of the KT and BIC estimators

For calculating the KT estimator we use the CTM algorithm of Willems, Shtarkov and Tjalkens (1993, 2000), and for the computation of the BIC estimator we use a similar algorithm.

Both algorithms have the following construction. Consider the full tree A^D, where $D = D(n) = o(\log n)$, and let $\mathcal{S}_D$ denote the set of its nodes, i.e., the set of all strings of length at most D. Based on the sample x_1^n we assign to each node a value and a binary indicator. This assignment is recursive, that is, the value and the indicator assigned to a node are calculated from the value assigned to the children of this node. The desired estimator will be the subtree determined by the indicators as specified below.

Consider the algorithm for the KT estimator. First, concentrate on the minimization of the criterion $\mathrm{KT}_{\mathcal{T}}(x_1^n)$ for $\mathcal{T} \in \mathcal{F}_1(x_1^n) \cap \mathcal{I}$ with $d(\mathcal{T}) \le D(n)$ (needed in the case $d(\mathcal{T}_0) < \infty$). Let us factorize the KT-probability as

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \frac{1}{|A|^{\lceil \log n \rceil}} \prod_{s \in \mathcal{T}} \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) \tag{2.2a}$$

where

$$\widetilde{P}_{\mathrm{KT},\,s}(x_1^n) = \begin{cases} \dfrac{\prod_{a:N_n(s,a)\ge 1} \left[\left(N_n(s,a) - \frac{1}{2}\right) \left(N_n(s,a) - \frac{3}{2}\right) \cdots \left(\frac{1}{2}\right) \right]}{\left(N_n(s) - 1 + \frac{|A|}{2}\right) \left(N_n(s) - 2 + \frac{|A|}{2}\right) \cdots \left(\frac{|A|}{2}\right)} & \text{if } N_n(s) \ge 1, \\ 1 & \text{if } N_n(s) = 0. \end{cases} \tag{2.2b}$$

Thus, the KT estimator can be written as

$$\begin{aligned}\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) &= \arg \min_{\mathcal{T}\in\mathcal{F}_1(x_1^n)\cap\mathcal{I},\, d(\mathcal{T})\le D(n)} \mathrm{KT}_{\mathcal{T}}(x_1^n) \\ &= \arg \max_{\mathcal{T}\in\mathcal{F}_1(x_1^n)\cap\mathcal{I},\, d(\mathcal{T})\le D(n)} \prod_{s\in\mathcal{T}} \widetilde{P}_{\mathrm{KT},\,s}(x_1^n).\end{aligned}$$

Definition 2.14. *Given a sample x_1^n, to each string $s \in \mathcal{S}_D$, $D = D(n)$ we assign recursively, starting from the leaves of the full tree A^D, the* KT-maximizing value *(no longer a probability)*

$$V_{D,\,s}^{\mathrm{KT}}(x_1^n) = \begin{cases} \max\left\{ \widetilde{P}_{\mathrm{KT},\,s}(x_1^n)\,,\ \prod_{a\in A} V_{D,\,as}^{\mathrm{KT}}(x_1^n) \right\} & \text{for } s\in\mathcal{S}_D,\ 0\le l(s)<D, \\ \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) & \text{for } s\in\mathcal{S}_D,\ l(s)=D, \end{cases}$$

and the KT-maximizing indicator

$$\chi_{D,\,s}^{\mathrm{KT}}(x_1^n) = \begin{cases} 1 & \text{if } \prod_{a\in A} V_{D,\,as}^{\mathrm{KT}}(x_1^n) > \widetilde{P}_{\mathrm{KT},\,s}(x_1^n)\ ;\ \text{for } s\in\mathcal{S}_D,\ 0\le l(s)<D, \\ 0 & \text{if } \prod_{a\in A} V_{D,\,as}^{\mathrm{KT}}(x_1^n) \le \widetilde{P}_{\mathrm{KT},\,s}(x_1^n)\ ;\ \text{for } s\in\mathcal{S}_D,\ 0\le l(s)<D, \\ 0 & \text{for } s\in\mathcal{S}_D,\ l(s)=D. \end{cases}$$

Using the KT-maximizing indicators, we assign to each $s \in \mathcal{S}_D$, $D = D(n)$ a KT-maximizing tree $\mathcal{T}_{D,\,s}^{\mathrm{KT}}(x_1^n) \succeq \{s\}$. This tree is defined recursively:

Definition 2.15. *Given $s \in \mathcal{S}_D$, start with $\mathcal{T} = \{s\}$ at step 0. At each step consider all s_1 in the tree $\mathcal{T}$ at that step whose indicator is 1, and the shortest $s_2 \succeq s_1$ such that there exist at least 2 letters $a \in A$ with $N_n(as_2) > 0$. Replace each s_1 by the set of its continuations as_2, $a \in A$ satisfying $N_n(as_2) > 0$, this yields the tree $\mathcal{T}$ for the next step. $\mathcal{T}_{D,\,s}^{\mathrm{KT}}(x_1^n)$ is defined as the tree $\mathcal{T}$ when this procedure stops.*

For this definition to be meaningful, it should be verified that to each $s_1 \in \mathcal{S}_D$ with indicator 1 there exists $s_2 \in \mathcal{S}_{D-1}$ with the properties in Definition 2.15. This follows from the facts that *(i)* $\chi_{D,\,s}^{\mathrm{KT}}(x_1^n) = 0$ if $l(s) = D$, and *(ii)* if $N_n(s) = N_n(as)$ holds for a string s and a letter a (and thus $N_n(a_1 s) = 0$ for all $a_1 \neq a$, $a_1 \in A$) then $\chi_{D,\,as}^{\mathrm{KT}}(x_1^n) = 0$ implies $\chi_{D,\,s}^{\mathrm{KT}}(x_1^n) = 0$.

Proposition 2.16. *The KT estimator equals the KT-maximizing tree assigned to the root, that is,*

$$\widehat{\mathcal{T}}_{\mathrm{KT}}(x_1^n) = \mathcal{T}_{D,\varnothing}^{\mathrm{KT}}(x_1^n).$$

Proof. The claimed equality is a consequence of $\mathcal{T}^{\mathrm{KT}}_{D,\varnothing}(x_1^n) \in \mathcal{F}_1(x_1^n) \cap \mathcal{I}$ and of the special case $s = \varnothing$ of the next lemma. □

For any $s \in \mathcal{S}_D$, define $\mathcal{F}_1(x_1^n|s)$ as the family of all trees $\mathcal{T}$ such that $N_n(us) \geq 1$ for all $u \in \mathcal{T}$, and each $s' \succ s$ with $N_n(s') \geq 1$ is either a postfix of us for some $u \in \mathcal{T}$ or has a postfix us with $u \in \mathcal{T}$.

Lemma 2.17. *For any $s \in \mathcal{S}_D$*

$$V^{\mathrm{KT}}_{D,\,s}(x_1^n) = \max_{\mathcal{T} \in \mathcal{F}_1(x_1^n|s):\, d(\mathcal{T}) \leq D - l(s)} \prod_{u \in \mathcal{T}} \widetilde{P}_{\mathrm{KT},\,us}(x_1^n) = \prod_{u \in \mathcal{T}^{\mathrm{KT}}_{D,\,s}(x_1^n)} \widetilde{P}_{\mathrm{KT},\,u}(x_1^n).$$

Proof. By induction on the length of the string s, similarly to (Willems, Shtarkov and Tjalkens, 1993). For $l(s) = D$ the statement is obvious.

Supposing the assertion holds for all strings of length d, for any s with $l(s) = d - 1$ we have

$$\begin{aligned}
V^{\mathrm{KT}}_{D,\,s}(x_1^n) &= \max \left\{ \widetilde{P}_{\mathrm{KT},\,s}(x_1^n)\,,\ \prod_{a \in A} V^{\mathrm{KT}}_{D,\,as}(x_1^n) \right\} \\
&= \max \left\{ \widetilde{P}_{\mathrm{KT},\,s}(x_1^n),\ \prod_{a \in A,\, N_n(as) \geq 1} \left(\max_{\mathcal{T}_a \in \mathcal{F}_1(x_1^n|as):\, d(\mathcal{T}_a) \leq D-d} \prod_{v \in \mathcal{T}_a} \widetilde{P}_{\mathrm{KT},\,vas}(x_1^n) \right) \right\} \\
&= \max \left\{ \widetilde{P}_{\mathrm{KT},\,s}(x_1^n)\,,\ \max_{\mathcal{T} \in \mathcal{F}_1(x_1^n|s):\, 1 \leq d(\mathcal{T}) \leq D-(d-1)} \prod_{u \in \mathcal{T}} \widetilde{P}_{\mathrm{KT},\,us}(x_1^n) \right\} \\
&= \max_{\mathcal{T} \in \mathcal{F}_1(x_1^n|s):\, d(\mathcal{T}) \leq D-(d-1)} \prod_{u \in \mathcal{T}} \widetilde{P}_{\mathrm{KT},\,us}(x_1^n).
\end{aligned}$$

Here the first equality holds by Definition 2.14, and the second one by the induction hypothesis, using the obvious fact that $V^{\mathrm{KT}}_{D,\,as}(x_1^n) = 1$ if $N_n(as) = 0$. The third equality follows since any family of trees $\mathcal{T}_a$, $a \in A$, $N_n(as) \geq 1$, satisfying the indicated constraints, uniquely corresponds to a tree $\mathcal{T} \in \mathcal{F}_1(x_1^n|s)$ with $1 \leq d(\mathcal{T}) \leq D-(d-1)$ via $\mathcal{T} = \cup_a \{\, sa : s \in \mathcal{T}_a \,\}$, and the last equality is obvious.

Moreover, due to Definitions 2.14 and 2.15, the induction hypothesis implies that the above maximum is attained for $\mathcal{T} = \mathcal{T}^{\mathrm{KT}}_{D,\,s}(x_1^n)$, proving the second equality of the assertion. □

Remark 2.18. Lemma 2.17 above, with the condition $\mathcal{T} \in \mathcal{F}_1(x_1^n|s)$ replaced by the condition that $\mathcal{T}$ is complete, is a result of Willems, Shtarkov and Tjalkens (1993, 2000) (with the minor difference that the trees there also had "costs"), and the above proof is similar to theirs. It follows, in particular, that the KT-maximizing values for complete trees are the same as for trees in the families

$\mathcal{F}_1(x_1^n|s)$. The KT-maximizing complete tree assigned to the root could be obtained from our $\mathcal{T}^{\mathrm{KT}}_{D,\varnothing}(x_1^n)$ by adding edges that did not occur in the sample. This no longer holds for the cases treated below. □

In the general case when the criterion $\mathrm{KT}_{\mathcal{T}}(x_1^n)$ has to be minimized for $\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I}$ with $d(\mathcal{T}) \le D(n)$, Proposition 2.16 still holds, with the same proof, if $\mathcal{F}_1(x_1^n|s)$ is replaced by $\mathcal{F}_{n^\alpha}(x_1^n|s)$ defined analogously, and Definition 2.15 of the KT-maximizing subtree is replaced by the following

Definition 2.19. *Given $s \in \mathcal{S}_D$, start with $\mathcal{T} = \{s\}$ at step 0. At each step consider all s_1 in the tree $\mathcal{T}$ at that step whose indicator is 1, and the shortest $s_2 \succeq s_1$ such that there exist at least 2 letters $a \in A$ with $N_n(as_2) > 0$. Replace those s_1 as above whose continuations as_2, $a \in A$ with $N_n(as_2) > 0$ all satisfy $N_n(as_2) \ge n^\alpha$, by the sets of these continuations. This yields the tree $\mathcal{T}$ for the next step. $\mathcal{T}^{\mathrm{KT}}_{D,\,s}(x_1^n)$ is defined as the tree $\mathcal{T}$ when this procedure stops.*

Consider next the algorithm for the BIC estimator. First, concentrate on the minimization of the criterion $\mathrm{BIC}_{\mathcal{T}}(x_1^n)$ for $\mathcal{T} \in \mathcal{F}_1(x_1^n) \cap \mathcal{I}$ with $d(\mathcal{T}) \le D(n)$ (needed in the case $d(\mathcal{T}_0) < \infty$). Factorize the maximum likelihood as

$$\mathrm{ML}_{\mathcal{T}}(x_1^n) = \prod_{s \in \mathcal{T}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) \tag{2.3a}$$

where

$$\widetilde{P}_{\mathrm{ML},\,s}(x_1^n) = \begin{cases} \displaystyle\prod_{a \in A} \left[\frac{N_n(s,a)}{N_n(s)} \right]^{N_n(s,a)} & \text{if } N_n(s) \ge 1, \\ 1 & \text{if } N_n(s) = 0. \end{cases} \tag{2.3b}$$

It will be convenient to consider the BIC estimator replacing n in the penalty term, see Definition 2.4, by a temporarily unspecified $\breve{n}$. In the proof of Theorem 2.12, $\breve{n} = n$ will be taken, and in the proof of Theorem 2.13, $\breve{n}$ will be chosen as the number replacing n in the statement of that theorem. Then

$$\begin{aligned} \widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) &= \arg \min_{\mathcal{T} \in \mathcal{F}_1(x_1^n) \cap \mathcal{I},\, d(\mathcal{T}) \le D(n)} \mathrm{BIC}_{\mathcal{T}}(x_1^n) \\ &= \arg \max_{\mathcal{T} \in \mathcal{F}_1(x_1^n) \cap \mathcal{I},\, d(\mathcal{T}) \le D(n)} \breve{n}^{-\frac{|A|-1}{2}|\mathcal{T}|} \prod_{s \in \mathcal{T}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n). \end{aligned}$$

The appropriate formalization of the algorithm is the following.

Definition 2.20. *Given a sample x_1^n, to each string $s \in \mathcal{S}_D$, $D = D(n)$ we assign recursively, starting from the leaves of the full tree A^D, the* BIC-maximizing value

$$V_{D,\,s}^{\mathrm{BIC}}(x_1^n) = \begin{cases} \max\left\{ \breve{n}^{-\frac{|A|-1}{2}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n)\,, \prod_{a\in A} V_{D,\,as}^{\mathrm{BIC}}(x_1^n) \right\} & \text{for } s \in \mathcal{S}_D,\ 0 \le l(s) < D, \\ \breve{n}^{-\frac{|A|-1}{2}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) & \text{for } s \in \mathcal{S}_D,\ l(s) = D, \end{cases}$$

and the BIC-maximizing indicator

$$\chi_{D,\,s}^{\mathrm{BIC}}(x_1^n) = \begin{cases} 1 \quad \text{if } \prod_{a\in A} V_{D,\,as}^{\mathrm{BIC}}(x_1^n) > \breve{n}^{-\frac{|A|-1}{2}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) \ ;\ \text{for } s \in \mathcal{S}_D,\ 0 \le l(s) < D, \\ 0 \quad \text{if } \prod_{a\in A} V_{D,\,as}^{\mathrm{BIC}}(x_1^n) \le \breve{n}^{-\frac{|A|-1}{2}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) \ ;\ \text{for } s \in \mathcal{S}_D,\ 0 \le l(s) < D, \\ 0 \quad \text{for } s \in \mathcal{S}_D,\ l(s) = D. \end{cases}$$

Using the BIC-maximizing indicators, we assign to each $s \in \mathcal{S}_D$, $D = D(n)$ a BIC-maximizing tree $\mathcal{T}_{D,\,s}^{\mathrm{BIC}}(x_1^n) \succeq \{s\}$.

Definition 2.21. *$\mathcal{T}_{D,\,s}^{\mathrm{BIC}}(x_1^n)$ is defined similarly as $\mathcal{T}_{D,\,s}^{\mathrm{KT}}(x_1^n)$ in Definition 2.15.*

Proposition 2.22. *The BIC estimator equals the BIC-maximizing tree assigned to the root, that is,*

$$\widehat{\mathcal{T}}_{\mathrm{BIC}}(x_1^n) = \mathcal{T}_{D,\varnothing}^{\mathrm{BIC}}(x_1^n).$$

Proof. The proof is similar to the KT case, the claimed equality is now a consequence of the special case $s = \varnothing$ of the lemma below. □

Lemma 2.23. *For any $s \in \mathcal{S}_D$*

$$\begin{aligned} V_{D,\,s}^{\mathrm{BIC}}(x_1^n) &= \max_{\mathcal{T}\in\mathcal{F}_1(x_1^n|s):\, d(\mathcal{T})\le D-l(s)} \breve{n}^{-\frac{|A|-1}{2}|\mathcal{T}|} \prod_{u\in\mathcal{T}} \widetilde{P}_{\mathrm{ML},\,us}(x_1^n) \\ &= \prod_{u\in\mathcal{T}_{D,\,s}^{\mathrm{BIC}}(x_1^n)} \breve{n}^{-\frac{|A|-1}{2}|\mathcal{T}_{D,\,s}^{\mathrm{BIC}}|} \widetilde{P}_{\mathrm{ML},\,us}(x_1^n). \end{aligned}$$

Proof. Analogous to the proof of Lemma 2.17:

$$V^{\mathrm{BIC}}_{D,\,s}(x_1^n) = \max\left\{ \breve{n}^{-\frac{|A|-1}{2}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n)\,,\ \prod_{a\in A} V^{\mathrm{BIC}}_{D,\,as}(x_1^n) \right\}$$

$$= \max\left\{ \breve{n}^{-\frac{|A|-1}{2}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n)\,, \right.$$

$$\left. \prod_{a\in A,\,N_n(as)\ge 1} \left(\max_{\mathcal{T}_a\in\mathcal{F}_1(x_1^n|as):\,d(\mathcal{T}_a)\le D-d} \breve{n}^{-\frac{|A|-1}{2}|\mathcal{T}_a|} \prod_{v\in\mathcal{T}_a} \widetilde{P}_{\mathrm{ML},\,vas}(x_1^n) \right) \right\}$$

$$= \max\left\{ \breve{n}^{-\frac{|A|-1}{2}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n)\,,\ \max_{\mathcal{T}\in\mathcal{F}_1(x_1^n|s):\,1\le d(\mathcal{T})\le D-(d-1)} \breve{n}^{-\frac{|A|-1}{2}|\mathcal{T}|} \prod_{u\in\mathcal{T}} \widetilde{P}_{\mathrm{ML},\,us}(x_1^n) \right\}$$

$$= \max_{\mathcal{T}\in\mathcal{F}_1(x_1^n|s):\,d(\mathcal{T})\le D-(d-1)} \breve{n}^{-\frac{|A|-1}{2}|\mathcal{T}|} \prod_{u\in\mathcal{T}} \widetilde{P}_{\mathrm{ML},\,us}(x_1^n),$$

where, for the third equality, we used that for a family of trees $\mathcal{T}_a$, $a \in A$, $N_n(as) \ge 1$ and the corresponding $\mathcal{T}$ as in the proof of Lemma 2.17, we have $|\mathcal{T}| = \sum_{a\in A,\,N_n(as)\ge 1} |\mathcal{T}_a|$. □

In the general case when the criterion $\mathrm{BIC}_{\mathcal{T}}(x_1^n)$ has to be minimized for $\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I}$ with $d(\mathcal{T}) \le D(n)$, Proposition 2.22 still holds, with the same proof, if $\mathcal{F}_1(x_1^n|s)$ is replaced by $\mathcal{F}_{n^\alpha}(x_1^n|s)$ defined analogously, and Definition 2.21 of the BIC-maximizing tree is replaced by the following

Definition 2.24. *$\mathcal{T}^{\mathrm{BIC}}_{D,\,s}(x_1^n)$ is defined similarly as $\mathcal{T}^{\mathrm{KT}}_{D,\,s}(x_1^n)$ in Definition 2.19.*

Finally, we show that the above algorithms have the asserted computational complexity in off-line and on-line cases.

Proof of Theorem 2.12. Since $D(n) = o(\log n)$, we may write $D(n) = \varepsilon_n \log n$, where $\varepsilon_n \to 0$. Denote $\widetilde{P}_s(x_1^n)$, $V_{D,\,s}(x_1^n)$, $\chi_{D,\,s}(x_1^n)$ either the values $\widetilde{P}_{\mathrm{KT},\,s}(x_1^n)$, $V^{\mathrm{KT}}_{D,\,s}(x_1^n)$, $\chi^{\mathrm{KT}}_{D,\,s}(x_1^n)$ or $\widetilde{P}_{\mathrm{ML},\,s}(x_1^n)$, $V^{\mathrm{BIC}}_{D,\,s}(x_1^n)$, $\chi^{\mathrm{BIC}}_{D,\,s}(x_1^n)$. In the BIC case we use $\breve{n} = n$ in Definition 2.20.

For each string $s \in \mathcal{S}_D$, $D = D(n) = \varepsilon_n \log n$, the counts $N_n(s,a)$, $a \in A$, and the values $\widetilde{P}_s(x_1^n)$, $V_{D,\,s}(x_1^n)$, $\chi_{D,\,s}(x_1^n)$ are stored. The number of stored data is proportional to the cardinality of $\mathcal{S}_D$, which is

$$\sum_{j=0}^{D} |A|^j = \frac{|A|^{D+1}-1}{|A|-1} \le 2|A|^D = O(n^\varepsilon). \tag{2.4}$$

To get the maximizing indicators $\chi_{D,\,s}(x_1^n)$, $s \in \mathcal{S}_D$ which give rise to the estimated context tree according to Definitions 2.15, 2.19, 2.21, 2.24, first we need the counts $N_n(s,a)$, $s \in \mathcal{S}_D$, $a \in A$.

The counts $N_n(s,a)$, $l(s)=D$, $a\in A$ can be determined successively processing the sample x_1^n from position $j=\lceil \log n\rceil$ to $j=n$, and at instance j incrementing the count $N_n\left(x_{j-D(n)}^{j-1}, x_j\right)$ by 1 (the starting values of all counts being 0). This is $O(n)$ calculations. The other counts $N_n(s,a)$, $s\in\mathcal{S}_{D-1}$, $a\in A$ can be determined recursively, as $N_n(s,a)=\sum_{b\in A} N_n(bs,a)$. This is $|A|\,|\mathcal{S}_{D-1}|=o(n)$ calculations.

Then, from these counts the values $\widetilde{P}_s(x_1^n)$ are determined by $O(n)$ multiplications. The calculation of the values $V_{D,\,s}(x_1^n)$ and $\chi_{D,\,s}(x_1^n)$ requires calculations proportional to the cardinality of $\mathcal{S}_D$, which is less than $2|A|^D=o(n)$. □

Proof of Theorem 2.13. We use the same notation as in the previous proof, except that in the BIC case $\breve{n}$ in Definition 2.20 is set equal to the smallest $k_m>n$ ($m\in\mathbb{N}$), see the statement of Theorem 2.13. Clearly, see (2.6) in the next section, for the BIC estimator with the increased penalty term in Theorem 2.13, the consistency assertions in Theorem 2.6 continue to hold.

The calculations required by the algorithms in Definitions 2.14 and 2.20 can be performed recursively in the sample size n.

Suppose at instant i, for each string $s\in\mathcal{S}_{D(i)}$, the counts $N_i(s,a)$, $a\in A$, and the values $\widetilde{P}_s(x_1^i)$, $V_{D,\,s}(x_1^i)$, $\chi_{D,\,s}(x_1^i)$ are stored. The number of stored data is proportional to the cardinality of $\mathcal{S}_{D(i)}$, which is $O(i^{\varepsilon})$, see (2.4).

Consider first those instances i when the sample size increases from $i-1$ to i but the depth does not change, $D(i)=D(i-1)$. If $\widetilde{P}_s(x_1^{i-1})$ at a node s is known, the value $\widetilde{P}_s(x_1^i)$ can be calculated using, for the KT case, that

$$\widetilde{P}_{\mathrm{KT},\,s}(x_1^i)=\frac{N_i(s,x_i)+1/2}{N_i(s)+|A|/2}\;\widetilde{P}_{\mathrm{KT},\,s}(x_1^{i-1}),$$

and in the BIC case that in the expression of $\widetilde{P}_{\mathrm{ML},\,s}(x_1^{i-1})$ only the counts $N_i(s,x_i)$ and $N_i(s)$ were incremented to obtain $\widetilde{P}_{\mathrm{ML},\,s}(x_1^i)$. From $\widetilde{P}_s(x_1^i)$ the values $V_{D,\,s}(x_1^i)$ and $\chi_{D,\,s}(x_1^i)$ can be computed in constant number of steps. These values are different for x_1^{i-1} and x_1^i only when s is a postfix of x_1^{i-1}, hence updating is needed at $D(i)$ nodes only. Thus the number of required computations is proportional to $D(i)$.

Consider next those instances i when the depth increases, $D(i)=D(i-1)+1$. In this case we have three tasks. We have to update the values $\widetilde{P}_s(x_1^{i-1})$ at those nodes s that already existed at instance $i-1$, namely where $l(s)<D(i)$. In addition, we have to calculate the values $\widetilde{P}_s(x_1^i)$ for the new terminal nodes s, $l(s)=D(i)$, and recalculate $V_{D,\,s}(x_1^i)$ and $\chi_{D,\,s}(x_1^i)$ at all nodes s of the new full tree. The former needs $O(i)$ calculations. Indeed, the counts $N_i(s,a)$, $l(s)=D(i)$

can be determined successively processing the sample x_1^i from position $j = \lceil \log n \rceil$ to $j = i$, and at instance j incrementing the count $N_i\left(x_{j-D(i)}^{j-1}, x_j\right)$ by 1 (the starting values of all counts being 0), and from these counts the values $\widetilde{P}_s(x_1^i)$ are determined by $O(i)$ multiplications. The recalculation of the values $V_{D,\,s}(x_1^i)$ and $\chi_{D,\,s}(x_1^i)$ requires calculations proportional to the cardinality of $\mathcal{S}_{D(i)}$, which is $O(i^\varepsilon)$.

Finally, the total number of computations performed on a sample x_1^n is bounded as follows. The number of computations needed for the updating at all instances $i \le n$ is proportional to

$$\sum_{i=1}^{n} D(i) = \sum_{i=1}^{n} \lfloor \varepsilon_i \log i \rfloor = o(n \log n).$$

The number of computations to calculate the values $\widetilde{P}_s$ for the new terminal nodes at the instances when $D(i)$ increases is proportional to

$$\sum_{D=0}^{\lfloor \varepsilon_n \log n \rfloor} \min\{\, i : D \le \varepsilon_i \log i \,\} = \sum_{D=0}^{\lfloor \varepsilon_n \log n \rfloor} \min\{\, i : 2^{D/\varepsilon_i} \le i \,\}$$
$$\le \sum_{D=0}^{\lfloor \varepsilon_n \log n \rfloor} 2^{D/\varepsilon_n} + 1 \le O\left(2^{\frac{1}{\varepsilon_n}\varepsilon_n \log n}\right) + \varepsilon_n \log n = O(n).$$

The number of computations to recalculate $V_{D,\,s}$, $\chi_{D,\,s}$ for all nodes in the full tree $A^{D(i)}$ at the instances when $D(i)$ increases is of order

$$\sum_{D=0}^{\lfloor \varepsilon_n \log n \rfloor} 2|A|^D = O\left(|A|^{\varepsilon_n \log n}\right) = o(n).$$

□

2.4 Consistency of the KT and BIC estimators

Proof of Theorem 2.6. In the case $d(\mathcal{T}_0) < \infty$, the properties $\mathcal{T} \in \mathcal{F}_1(x_1^n)$, $d(\mathcal{T}) \le D(n)$, $D(n) = o(\log n)$, imply that $\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n)$, eventually almost surely as $n \to \infty$, by Lemma 2.27 in the Appendix. Hence it suffices to prove the second assertion of the Theorem.

The proof is indirect. Fix an arbitrary constant K. It suffices to show that if $\mathcal{T}\big|_K \ne \mathcal{T}_0\big|_K$ for some $\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I}$ with $d(\mathcal{T}) \le D(n)$, $D(n) = o(\log n)$, then there exists a modification $\mathcal{T}'$ of $\mathcal{T}$ also satisfying $\mathcal{T}' \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I}$, $d(\mathcal{T}') \le D(n)$ such that

$$\mathrm{BIC}_{\mathcal{T}}(x_1^n) > \mathrm{BIC}_{\mathcal{T}'}(x_1^n), \tag{2.5}$$

simultaneously for all considered trees $\mathcal{T}$, eventually almost surely as $n \to \infty$.

Recall the factorization (2.3a):

$$\mathrm{ML}_{\mathcal{T}}(x_1^n) = \prod_{s\in\mathcal{T}} \widetilde{P}_{\mathrm{ML},\,s}(x_1^n).$$

Using this and the definition of BIC, see Definition 2.4, (2.5) is equivalent to

$$\sum_{s\in\mathcal{T}} \log \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) - \sum_{s'\in\mathcal{T}'} \log \widetilde{P}_{\mathrm{ML},\,s'}(x_1^n) < \frac{(|A|-1)}{2}\,(|\mathcal{T}| - |\mathcal{T}'|)\log n. \tag{2.6}$$

Since $\mathcal{T}$ is a feasible tree by assumption, so is also $\mathcal{T}\big|_K$ defined by (2.1). For n sufficiently large, so that $N_n(s) \geq 1$ for all s with $l(s) \leq K$, $Q(s) > 0$, it follows by Remark 2.2 that $\mathcal{T}_0\big|_K$ is feasible, as well. Hence, the indirect assumption $\mathcal{T}\big|_K \neq \mathcal{T}_0\big|_K$ implies that there exist strings $\tilde{s} \in \mathcal{T}\big|_K$ and $\tilde{s_0} \in \mathcal{T}_0\big|_K$ such that either $\tilde{s} \prec \tilde{s_0}$ (underestimation) or $\tilde{s} \succ \tilde{s_0}$ (overestimation). Equivalently, there exist $s \in \mathcal{T}$ and $s_0 \in \mathcal{T}_0$ such that either (a) $l(s) < K$, $s \prec s_0$ or (b) $l(s_0) < K$, $s_0 \prec s$.

In case (a), apply Lemma 2.25 below with $\tilde{s} = s$. For $\widetilde{\mathcal{T}}$ in Lemma 2.25 we have

$$\mathcal{T}' = (\,\mathcal{T}\backslash\{s\}\,) \cup \widetilde{\mathcal{T}} \;\; \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I},$$

and $\mathcal{T}'$ satisfies (2.6), since the right hand side of (2.7) is smaller than $\nu(1 - |\widetilde{\mathcal{T}}|)\log n$, for any $\nu > 0$, if n is sufficiently large. Moreover, (2.6) holds simultaneously for all considered $\mathcal{T}$, since the number of possible strings $s \in \mathcal{T}$, $l(s) \leq K$ is finite.

In case (b), apply Lemma 2.26 below with $s' = s$. Then with $\widetilde{\mathcal{T}}$ and w as in Lemma 2.26 we have

$$\mathcal{T}' = (\,\mathcal{T}\backslash\widetilde{\mathcal{T}}\,) \cup \{w\} \;\; \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I},$$

and this $\mathcal{T}'$ satisfies (2.6) simultaneously for all considered $\mathcal{T}$, eventually almost surely as $n \to \infty$. □

Lemma 2.25. *For any proper postfix $\tilde{s}$ of some $s_0 \in \mathcal{T}_0$, there exists an irreducible tree $\widetilde{\mathcal{T}} \succ \{\tilde{s}\}$ with $d(\widetilde{\mathcal{T}}) < \infty$ such that $N_n(s) \geq n^\alpha$ for each $s \in \widetilde{\mathcal{T}}$, each $v \succeq \tilde{s}$ with $N_n(v) \geq 1$ has a postfix in $\widetilde{\mathcal{T}}$, and*

$$\log \widetilde{P}_{\mathrm{ML},\,\tilde{s}}(x_1^n) - \sum_{s\in\widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) < -c\,n, \tag{2.7}$$

eventually almost surely as $n \to \infty$, where $c > 0$ is a sufficiently small constant.

Proof. Given $\tilde{s} \prec s_0 \in \mathcal{T}_0$, denote by s_{0l} the l-length postfix of s_0. Let

$$\widetilde{\mathcal{T}} = \{\, as_{0l} : l(\tilde{s}) \leq l \leq L,\, a \in A,\, as_{0l} \neq s_{0l+1},\, Q(as_{0l}) > 0 \,\} \cup \{\, s_{0L+1} \,\}.$$

We show that if $L = l(s_0) - 1$ when $l(s_0) < \infty$, or L is sufficiently large with the property $Q(s_{0L+1}) < Q(s_{0L})$ when $l(s_0) = \infty$, this tree $\widetilde{\mathcal{T}}$ satisfies the assertions of the Lemma.

Now, using (2.3b), the inequality (2.7) can be written as

$$\sum_{s\in\widetilde{\mathcal{T}},\, a\in A} N_n(s,a) \log \frac{N_n(s,a)}{N_n(s)} - \sum_{a\in A} N_n(\tilde{s},a) \log \frac{N_n(\tilde{s},a)}{N_n(\tilde{s})} > c\, n.$$

Due to the ergodic theorem, $N_n(v,a)/n \to Q(va)$ for any string v, almost surely as $n \to \infty$. Hence, it is enough to show that

$$\sum_{s\in\widetilde{\mathcal{T}},\, a\in A} Q(sa) \log \frac{Q(sa)}{Q(s)} - \sum_{a\in A} Q(\tilde{s}a) \log \frac{Q(\tilde{s}a)}{Q(\tilde{s})} > 0.$$

Jensen's inequality implies

$$Q(\tilde{s}) \sum_{s\in\widetilde{\mathcal{T}}} \frac{Q(s)}{Q(\tilde{s})} \left(\frac{Q(sa)}{Q(s)} \log \frac{Q(sa)}{Q(s)} \right) \geq Q(\tilde{s}a) \log \frac{Q(\tilde{s}a)}{Q(\tilde{s})}, \quad a \in A,$$

where the strict inequality holds for some $a \in A$, unless $Q(a|\,\tilde{s}) = Q(a|\,s)$ for each $a \in A$ and $s \in \widetilde{\mathcal{T}}$, in particular, for $s = s_{0L+1}$. In the case $l(s_0) < \infty$ we have $s_{0L+1} = s_0$, hence the last contingency is ruled out by $\tilde{s} \prec s_0 \in \mathcal{T}_0$ and the definition of context tree $\mathcal{T}_0$. In the case $l(s_0) = \infty$, if $Q(a|\tilde{s})$ were equal to $Q(a|s_{0L+1})$ for each $a \in A$ and all L satisfying $Q(s_{0L+1}) < Q(s_{0L})$, letting $L \to \infty$ would give $Q(a|\tilde{s}) = Q(a|s_0)$, again contradicting $\tilde{s} \prec s_0 \in \mathcal{T}_0$.

The irreducibility of $\widetilde{\mathcal{T}}$ is obvious when $l(s_0) = \infty$, and in the case $l(s_0) < \infty$ it only requires checking that for $L = l(s_0) - 1$ there exists $a \in A$ with $as_{0L} \neq s_0$, $Q(as_{0L}) > 0$; this follows from $s_0 \in \mathcal{T}_0$ by Definition 2.1.

Moreover, we have $N_n(s) \geq n^\alpha$ for each $s \in \widetilde{\mathcal{T}}$ eventually almost surely as $n \to \infty$, on account of the ergodic theorem and $d(\widetilde{\mathcal{T}}) < \infty$. Finally, each $s \succeq \tilde{s}$ with $N_n(s) \geq 1$, in fact any $s \succeq \tilde{s}$ with $Q(s) > 0$, has a postfix in $\widetilde{\mathcal{T}}$ by its construction. □

Lemma 2.26. *For any tree $\mathcal{T} \in \mathcal{I}$ with $d(\mathcal{T}) \leq D(n)$, $D(n) = o(\log n)$, and $s' \in \mathcal{T}$ that has a proper postfix $s_0 \in \mathcal{T}_0$ with $l(s_0) \leq K$, there exists w satisfying $s' \succ w \succeq s_0$ such that, for $\widetilde{\mathcal{T}} = \{\, s \in \mathcal{T} : s \succ w \,\}$ and arbitrary $\nu > 0$,*

$$\sum_{s\in\widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) - \log \widetilde{P}_{\mathrm{ML},\,w}(x_1^n) < \nu(\,|\widetilde{\mathcal{T}}| - 1\,) \log n,$$

holds simultaneously for all considered $\mathcal{T}$ and s', eventually almost surely as $n \to \infty$. Moreover, here $w = a_{-k}a_{-k+1}\ldots a_{-1}$ can be chosen such that $a_{-k+1}\ldots a_{-1}$ is a proper postfix of some $s \in \mathcal{T}\backslash\widetilde{\mathcal{T}}$.

Proof. Let $w = a_{-k}a_{-k+1}\dots a_{-1}$ be the longest postfix of s' with $k < l(s')$ for which there exists a string in $\mathcal{T}$ not equal to w but having the postfix $a_{-k+1}\dots a_{-1}$. Then $\mathcal{T}_0 \in \mathcal{I}$ implies that $w \succeq s_0$, and hence $a_{-k+1}\dots a_{-1} \prec s$ for some $s \in \mathcal{T}\backslash\widetilde{\mathcal{T}}$.

Since

$$\prod_{a\in A}\left[\frac{N_n(w,a)}{N_n(w)}\right]^{N_n(w,a)} \geq \prod_{a\in A} Q(a|\,w)^{N_n(w,a)},$$

the left hand side of the claimed inequality can be bounded above by

$$\begin{aligned}
&\sum_{s\in\widetilde{\mathcal{T}},\,a\in A} N_n(s,a)\log\frac{N_n(s,a)}{N_n(s)} \;-\sum_{a\in A} N_n(w,a)\log Q(a|\,w)\\
&\quad\overset{(i)}{=}\sum_{s\in\widetilde{\mathcal{T}},\,a\in A} N_n(s,a)\log\frac{N_n(s,a)}{N_n(s)} \;-\sum_{s\in\widetilde{\mathcal{T}},\,a\in A} N_n(s,a)\log Q(a|\,s)\\
&\quad=\sum_{s\in\widetilde{\mathcal{T}}} N_n(s)\sum_{a\in A}\frac{N_n(s,a)}{N_n(s)}\log\frac{N_n(s,a)/N_n(s)}{Q(a|\,s)}\\
&\quad=\sum_{s\in\widetilde{\mathcal{T}}} N_n(s)\;D\left(\frac{N_n(s,\cdot)}{N_n(s)}\;\middle\|\;Q(\,\cdot\,|s)\right)
\end{aligned}$$

Here (i) follows as $s \succ w \succeq s_0 \in \mathcal{T}_0$ implies $Q(a|\,s) = Q(a|\,w) = Q(a|\,s_0)$ by Definition 2.1. Using Lemmas 2.28 and 2.29 in the Appendix, this can be further bounded above, eventually almost surely simultaneously for all considered $\mathcal{T}$ and s', by

$$\begin{aligned}
&\sum_{s\in\widetilde{\mathcal{T}}} N_n(s)\,\frac{1}{q_{\min}}\sum_{a\in A}\left[\frac{N_n(s,a)}{N_n(s)} - Q(a|\,s)\right]^2\\
&\qquad<\sum_{s\in\widetilde{\mathcal{T}}} N_n(s)\,\frac{1}{q_{\min}}\,|A|\,\frac{\delta\,\log N_n(s)}{N_n(s)} \leq \frac{\delta\,|A|}{q_{\min}}\,|\widetilde{\mathcal{T}}|\log n,
\end{aligned}$$

where $q_{\min}$ is the minimum of the nonzero conditional probabilities $Q(a|\,s_0)$, $a \in A$, $s_0 \in \mathcal{T}_0$, $l(s_0) \leq K$, and $\delta > 0$ is arbitrary small. □

Proof of Theorem 2.9. Similarly to the Proof of Theorem 2.6, it suffices to prove the second assertion.

The proof is indirect. Fix an arbitrary constant K. It suffices to show that if $\mathcal{T}\big|_K \neq \mathcal{T}_0\big|_K$ for some $\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I}$ with $d(\mathcal{T}) \leq D(n)$, $D(n) = o(\log n)$, then there exists a modification $\mathcal{T}'$ of $\mathcal{T}$ also satisfying $\mathcal{T}' \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I}$, $d(\mathcal{T}') \leq D(n)$ such that

$$\mathrm{KT}_{\mathcal{T}}(x_1^n) > \mathrm{KT}_{\mathcal{T}'}(x_1^n), \tag{2.8}$$

simultaneously for all considered trees $\mathcal{T}$, eventually almost surely as $n \to \infty$.

Recall the factorization (2.2a):

$$P_{\mathrm{KT},\mathcal{T}}(x_1^n) = \prod_{s\in\mathcal{T}} \widetilde{P}_{\mathrm{KT},\,s}(x_1^n).$$

Using this and the definition of KT, see Definition 2.8, (2.8) is equivalent to

$$\sum_{s'\in\mathcal{T}'} \log \widetilde{P}_{\mathrm{KT},\,s'}(x_1^n) - \sum_{s\in\mathcal{T}} \log \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) > 0. \tag{2.9}$$

First, for any string s with $N_n(s) \geq 1$ we use the standard bound (see, e.g., eq. (2.12) in (Csiszár and Shields, 2000))

$$\left| \log \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) - \sum_{a\in A} N_n(s,a) \log \frac{N_n(s,a)}{N_n(s)} + \frac{|A|-1}{2} \log N_n(s) \right| < C,$$

where C is a constant depending only on the alphabet size $|A|$. This implies for any tree $\mathcal{T} \in \mathcal{F}_1(x_1^n)$

$$\left| \sum_{s\in\mathcal{T}} \log \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) - \sum_{s\in\mathcal{T}} \log \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) + \frac{|A|-1}{2} \sum_{s\in\mathcal{T}} \log N_n(s) \right| < C|\mathcal{T}|\,. \tag{2.10}$$

Since $\mathcal{T} \in \mathcal{F}_{n^\alpha}(x_1^n)$ implies, by definition, $\log N_n(s) > \alpha \log n$ for $s \in \mathcal{T}$, it follows that

$$\sum_{s\in\mathcal{T}} \log \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) - \sum_{s\in\mathcal{T}} \log \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) > \frac{|A|-1}{2}\,|\mathcal{T}|\,\alpha \log n - C|\mathcal{T}|\,. \tag{2.11}$$

On the other hand, (2.10) implies the following upper bound for any $\mathcal{T} \in \mathcal{F}_1(x_1^n)$:

$$\sum_{s\in\mathcal{T}} \log \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) - \sum_{s\in\mathcal{T}} \log \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) < \frac{|A|-1}{2}\,|\mathcal{T}|\,\log n + C|\mathcal{T}|. \tag{2.12}$$

As in the proof of Theorem 2.6, the indirect assumption $\mathcal{T}\big|_K \neq \mathcal{T}_0\big|_K$ implies that there exist $s \in \mathcal{T}$ and $s_0 \in \mathcal{T}_0$ such that either (a) $l(s) < K$, $s \prec s_0$ or (b) $l(s_0) < K$, $s_0 \prec s$.

In case (a), apply Lemma 2.25 above with $\tilde{s} = s$. For $\widetilde{\mathcal{T}}$ in Lemma 2.25 we have

$$\begin{aligned}
\sum_{u\in\widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{KT},\,u}(x_1^n) - \log \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) &= \left(\sum_{u\in\widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{KT},\,u}(x_1^n) - \sum_{u\in\widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{ML},\,u}(x_1^n) \right) \\
+ \left(\sum_{u\in\widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{ML},\,u}(x_1^n) - \log \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) \right) &+ \left(\log \widetilde{P}_{\mathrm{ML},\,s}(x_1^n) - \log \widetilde{P}_{\mathrm{KT},\,s}(x_1^n) \right) \\
&\overset{(i)}{>} c\,n + \frac{|A|-1}{2}\,\alpha\,\log n - \frac{|A|-1}{2}\,|\widetilde{\mathcal{T}}|\,\log n - C\left(1 + |\widetilde{\mathcal{T}}|\right) > 0,
\end{aligned}$$

eventually almost surely as $n \to \infty$, where (i) follows using (2.11) with $\mathcal{T} = \{s\}$, (2.12) with $\mathcal{T} = \widetilde{\mathcal{T}}$, and Lemma 2.25. This gives that

$$\mathcal{T}' = (\,\mathcal{T} \backslash \{s\}\,) \cup \widetilde{\mathcal{T}} \;\; \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I}$$

satisfies (2.9), and (2.9) holds simultaneously for all considered $\mathcal{T}$, since the number of possible strings $s \in \mathcal{T}$, $l(s) \le K$ is finite.

In case (b), apply Lemma 2.26 above with $s' = s$. Then with $\widetilde{\mathcal{T}}$ and w as in Lemma 2.26 we have

$$\log \widetilde{P}_{\mathrm{KT},\,w}(x_1^n) - \sum_{u \in \widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{KT},\,u}(x_1^n) = \Big(\log \widetilde{P}_{\mathrm{KT},\,w}(x_1^n) - \log \widetilde{P}_{\mathrm{ML},\,w}(x_1^n)\Big)$$

$$+ \left(\log \widetilde{P}_{\mathrm{ML},\,w}(x_1^n) - \sum_{u \in \widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{ML},\,u}(x_1^n)\right) + \left(\sum_{u \in \widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{ML},\,u}(x_1^n) - \sum_{u \in \widetilde{\mathcal{T}}} \log \widetilde{P}_{\mathrm{KT},\,u}(x_1^n)\right)$$

$$\overset{(ii)}{>} \frac{|A|-1}{2}\Big(|\widetilde{\mathcal{T}}|(\alpha-\nu) - 1\Big)\log n - C(|\widetilde{\mathcal{T}}|+1) > 0,$$

eventually almost surely as $n \to \infty$, simultaneously for all considered $\mathcal{T}$, where (ii) follows using (2.11) with $\mathcal{T} = \widetilde{\mathcal{T}}$, (2.12) with $\mathcal{T} = \{w\}$, and Lemma 2.26. This gives that

$$\mathcal{T}' = (\,\mathcal{T} \backslash \widetilde{\mathcal{T}}\,) \cup \{w\} \;\; \in \mathcal{F}_{n^\alpha}(x_1^n) \cap \mathcal{I}$$

satisfies (2.9), eventually almost surely as $n \to \infty$, simultaneously for all considered $\mathcal{T}$. □

2.5 Discussion

We have proved the strong consistency of the BIC estimator and the KT version of MDL estimator of the context tree of any (stationary ergodic) process, when the depth of the hypothetical context trees is allowed to grow with the sample size n as $o(\log n)$. This context tree may have infinite depth, and it is not necessarily complete. These consistency results are generalizations of similar results for estimation of the order of Markov chains (Csiszár and Shields, 2000), (Csiszár, 2002).

We have considered process with time domain equal to the set of all integers, but as long as stationarity and ergodicity are insisted upon, any process with one-sided time domain $\mathbb{N}$ can be obtained by restricting the time domain of a process of the former kind. When dealing with Markov chain order estimation in the one-sided case, dropping the stationarity assumption causes no additional difficulty, see (Csiszár and Shields, 2000). For context tree estimation of tree

sources, non-stationarity may cause technical problems in dealing with transient phenomena, but does not appear to significantly change the picture, see (Martín, Seroussi and Weinberger, 2004).

While the BIC Markov order estimator is consistent without any bound on the hypothetical orders (Csiszár and Shields, 2000), it remains open whether the BIC context tree estimator remains consistent when dropping the depth bound $o(\log n)$, or replacing it by a bound $c \log n$. For the KT context tree estimator it also remains open whether the depth bound could be increased; it certainly can not be dropped or replaced by a large constant times $\log n$, since then consistency fails even for Markov order estimation (Csiszár and Shields, 2000).

Both with BIC and KT, we have considered two kinds of estimators, the second kind admitting only "r-frequent" hypothetical trees with $r = n^\alpha$. The latter conforms to the intuitive idea that the estimation should be based on those strings that "frequently" appeared in the sample, see (Bühlmann and Wyner, 1999). When the context tree has finite depth, the restriction to n^α-frequent hypothetical trees was not necessary since all feasible trees (of depth $D(n) = o(\log n)$) satisfied it automatically, eventually almost surely. It remains open whether the mentioned restriction is necessary for consistency when the context tree has infinite depth, and also whether the technical condition $\alpha > 1/2$ we need in the KT (but not in the BIC) case is really necessary.

A consequence of the consistency theorems is that when a process is not a Markov chain of any (finite) order, the estimated order, produced by either of the BIC or KT estimators tends to infinity almost surely.

The NML version of MDL was not considered for the context tree estimation problem (unlike for Markov order estimation (Csiszár, 2002)), because the structure of the NML criterion, unlike BIC and KT, appears unsuitable for CTM implementation.

We have also shown that the BIC and KT context tree estimators can be computed in linear time, via suitable modifications of the CTM method (Willems, Shtarkov and Tjalkens, 1993, 2000). An on-line procedure was also considered that calculates the estimators for all sample sizes $i \leq n$ in $o(n \log n)$ time. This result may be useful, for example, to implement context tree estimation with a stopping rule based on "stabilizing" of the estimator.

Finally we note that in the definition of BIC (Definition 2.4), the factor $(|A|-1)|\mathcal{T}|/2$ in the penalty term could be replaced by $c|\mathcal{T}|$, with any positive constant c, without affecting our results. The question of what other penalty terms might be appropriate is not in the scope of this work.

2.A Appendix

Lemma 2.27. *Given a process Q with context tree of finite depth, for any $0 < \alpha < 1$ there exists $\kappa > 0$ such that, eventually almost surely as $n \to \infty$,*

$$N_n(s) \geq n^\alpha,$$

simultaneously for all strings s with $l(s) \leq \kappa \log n$.

Proof. This bound has been used in (Csiszár, 2002), proof of Theorem 5. It is a consequence of the typicality theorem in (Csiszár and Shields, 2000), see also (Csiszár, 2002), remark after Th. 1. Indeed, the latter implies the existence of $\kappa > 0$ such that $N_n(s)/n \geq Q(s)/2$ simultaneously for all s with $l(s) < \kappa \log n$, eventually almost surely as $n \to \infty$. The assertion of the lemma follows, since $Q(s)$, when positive, is bounded below by $\xi^{l(s)}$ for a constant $\xi > 0$. □

Lemma 2.28. *Given a process Q and arbitrary $\alpha > 0$, $\delta > 0$, there exists $\kappa > 0$ such that, eventually almost surely as $n \to \infty$,*

$$\left| \frac{N_n(s,a)}{N_n(s)} - Q(a|\, s) \right| < \sqrt{\frac{\delta \, \log N_n(s)}{N_n(s)}}$$

simultaneously for all strings s with $l(s) \leq \kappa \log n$ and $N_n(s) \geq n^\alpha$ which have a postfix in the context tree of Q.

Proof. This is, in effect, Corollary 2 in (Csiszár, 2002). While that Corollary is stated for Markov processes only, the proof relies upon the martingale property of the sequence Z_n of (Csiszár, 2002), eq. (10). $Z_n = N_n(s,a) - Q(a|s)\, N_{n-1}(s)$ defines a martingale whenever s has a postfix in the context tree of Q, and the mentioned proof applies literally. □

Lemma 2.29. *If P_1 and P_2 are probability distributions on A satisfying*

$$\frac{1}{2}\, P_2(a) \leq P_1(a) \leq 2\, P_2(a), \quad a \in A,$$

then

$$D(P_1 \| P_2) \leq \sum_{a \in A} \frac{(P_1(a) - P_2(a))^2}{P_2(a)}.$$

Proof. See (Csiszár, 2002), Lemma 4. □

Chapter 3

Consistent Estimation of the Basic Neighborhood of Markov Random Fields

3.1 Introduction

In this chapter, Markov random fields on the lattice $\mathbb{Z}^d$ with finite state space are considered, adopting the usual assumption that the finite dimensional distributions are strictly positive. Equivalently, these are Gibbs fields with finite range interaction, see Georgii (1988). They are essential in statistical physics, for modeling interactive particle systems, Dobrushin (1968), and also in several other fields, Besag (1974), for example, in image processing, Azencott (1987).

One statistical problem for Markov random fields is parameter estimation when the interaction structure is known. By this we mean knowledge of the *basic neighborhood*, the minimal lattice region that determines the conditional distribution at a site on the condition that the values at all other sites are given; formal definitions are in Section 3.2. The conditional probabilities involved, assumed translation invariant, are parameters of the model. Note that they need not uniquely determine the joint distribution on $\mathbb{Z}^d$, a phenomenon known as *phase transition*. Another statistical problem is *model selection*, that is, the statistical estimation of the interaction structure (the basic neighborhood). This work is primarily devoted to the latter.

Parameter estimation for Markov random fields with a known interaction structure was considered, among others, by Pickard (1987), Gidas (1986), (1991), Geman and Graffigne (1987), Comets (1992). Typically, parameter estimation does not directly address the conditional probabilities mentioned above, but

rather the *potential*. This admits parsimonious representation of the conditional probabilities that are not free parameters, but have to satisfy algebraic conditions that need not concern us here. For our purposes, however, potentials will not be needed.

We are not aware of papers addressing model selection in the context of Markov random fields. In other contexts, penalized likelihood methods are popular, see Akaike (1972), Schwarz (1978). The Bayesian Information Criterion (BIC) of Schwarz (1978) has been proved to lead to consistent estimation of the "order of the model" in various cases, such as i.i.d. processes with distributions from exponential families, Haughton (1988), autoregressive processes, Hannan and Quinn (1979), and Markov chains, Finesso (1992). These proofs include the assumption that the number of candidate model classes is finite, for Markov chains this means that there is a known upper bound on the order of the process. The consistency of the BIC estimator of the order of a Markov chain without such prior bound was proved by Csiszár and Shields (2000); further related results appear in Csiszár (2002). A related recent result, for processes with variable memory length (Weinberger, Rissanen and Feder (1995), Bühlmann and Wyner (1999)), is the consistency of the BIC estimator of the context tree, without any prior bound on memory depth, Csiszár and Talata (2004).

For Markov random fields, penalized likelihood estimators like BIC run into the problem that the likelihood function can not be calculated explicitly. In addition, no simple formula is available for the "number of free parameters" typically used in the penalty term. To overcome these problems, we will replace likelihood by pseudo-likelihood, first introduced by Besag (1975), and modify also the penalty term; this will lead us to an analogue of BIC called the *Pseudo-Bayesian Information Criterion* or PIC. Our main result is that minimizing this criterion for a family of hypothetical basic neighborhoods that grows with the sample size at a specified rate, the resulting PIC estimate of the basic neighborhood equals the true one, eventually almost surely. In particular, the consistency theorem does not require a prior upper bound on the size of the basic neighborhood. It should be emphasized that the underlying Markov field need not be stationary (translation invariant), and phase transition causes no difficulty.

An auxiliary result perhaps of independent interest is a typicality proposition on the uniform closeness of empirical conditional probabilities to the true ones, for conditioning regions whose size may grow with the sample size. Though this result is weaker than analogous ones for Markov chains in Csiszár (2002), it will be sufficient for our purposes.

The structure of this chapter is the following. In Section 3.2 we introduce the basic notation and definitions, and formulate the main result. Its proof is provided by the propositions in Sections 3.4 and 3.5. Section 3.3 contains the statement and proof of the typicality proposition. Section 3.4 excludes overestimation, that is, the possibility that the estimated basic neighborhood properly contains the true one, using the typicality proposition. Section 3.5 excludes underestimation, that is, the possibility that the estimated basic neighborhood does not contain the true one, via an entropy argument and a modification of the typicality result. Section 3.6 is a discussion of the results. The Appendix contains some technical lemmas.

3.2 Notation and statement of the main results

We consider the d-dimensional *lattice* $\mathbb{Z}^d$. The points $i \in \mathbb{Z}^d$ are called sites, and $\|i\|$ denotes the maximum norm of i, that is, the maximum of the absolute values of the coordinates of i. The cardinality of a finite set Δ is denoted by $|\Delta|$. The notations $\subseteq$ and $\subset$ of inclusion and strict inclusion are distinguished in the dissertation.

A *random field* is a family of random variables indexed by the sites of the lattice, $\{ X(i) : i \in \mathbb{Z}^d \}$, where each $X(i)$ is a random variable with values in a finite set A. For $\Delta \subseteq \mathbb{Z}^d$, a region of the lattice, we write $X(\Delta) = \{ X(i) : i \in \Delta \}$. For the realizations of $X(\Delta)$ we use the notation $a(\Delta) = \{ a(i) \in A : i \in \Delta \}$. When Δ is finite, the $|\Delta|$-tuples $a(\Delta) \in A^\Delta$ will be referred to as *blocks*.

The joint distribution of the random variables $X(i)$ is denoted by Q. We assume that its finite dimensional marginals are strictly positive, that is,

$$Q(a(\Delta)) = \mathrm{Prob}\{ X(\Delta) = a(\Delta) \} > 0 \quad \text{for } \Delta \subset \mathbb{Z}^d \text{ finite, } a(\Delta) \in A^\Delta.$$

The last standard assumption admits unambiguous definition of the conditional probabilities

$$Q(a(\Delta) | \, a(\Phi)) = \mathrm{Prob}\{ X(\Delta) = a(\Delta) \mid X(\Phi) = a(\Phi) \}$$

for all disjoint finite regions Δ and Φ.

By a *neighborhood* Γ (of the origin 0) we mean a finite, central-symmetric set of sites with $0 \notin \Gamma$. Its radius is $r(\Gamma) = \max_{i \in \Gamma} \|i\|$. For any $\Delta \subseteq \mathbb{Z}^d$, its translate when 0 is translated to i is denoted by Δ^i. The translate Γ^i of a neighborhood Γ (of the origin) will be called the Γ-neighborhood of the site i, see Fig.3.1.

A *Markov random field* is a random field as above such that there exists a neighborhood Γ, called a *Markov neighborhood*, satisfying for every $i \in \mathbb{Z}^d$

$$Q(a(i)|\, a(\Delta^i)) = Q(a(i)|\, a(\Gamma^i)) \quad \text{if } \Delta \supset \Gamma,\ 0 \notin \Delta, \tag{3.1}$$

where the last conditional probability is translation invariant.

This concept is equivalent to that of a Gibbs field with a finite range interaction, see Georgii (1988). Motivated by this fact, the matrix

$$Q_\Gamma = \left\{\, Q_\Gamma(a|\, a(\Gamma)) : a \in A,\ a(\Gamma) \in A^\Gamma \,\right\}$$

specifying the (positive, translation invariant) conditional probabilities in (3.1) will be called *one-point specification*. All distributions on $A^{\mathbb{Z}^d}$ that satisfy (3.1) with a given conditional probability matrix Q_Γ are called *Gibbs distributions* with one-point specification Q_Γ. The distribution Q of the given Markov random field is one of these; Q is not necessarily translation invariant.

The following lemma summarizes some well-known facts; their formal derivation from results in Georgii (1988) is indicated in the Appendix.

Lemma 3.1. *For a Markov random field on the lattice as above, there exists a neighborhood Γ_0 such that the Markov neighborhoods are exactly those that contain Γ_0. Moreover, the global Markov property*

$$Q(a(\Delta)|\, a(\mathbb{Z}^d \setminus \Delta)) = Q(a(\Delta)|\, a(\cup_{i\in\Delta}\Gamma_0^i \setminus \Delta))$$

holds for each finite region $\Delta \subset \mathbb{Z}^d$. These conditional probabilities are translation invariant and uniquely determined by the one-point specification Q_{Γ_0}.

The smallest Markov neighborhood Γ_0 of Lemma 3.1 will be called the *basic neighborhood*. The minimal element of the corresponding one-point specification matrix Q_{Γ_0} is denoted by $q_{\min}$:

$$q_{\min} = \min_{a\in A,\, a(\Gamma_0)\in A^{\Gamma_0}} Q_{\Gamma_0}(a|\, a(\Gamma_0)) > 0.$$

In this chapter, we are concerned with the statistical estimation of the basic neighborhood Γ_0 from observing a realization of the Markov random field on an increasing sequence of finite regions $\Lambda_n \subset \mathbb{Z}^d$, $n \in \mathbb{N}$; thus the n'th sample is $x(\Lambda_n)$.

We will draw the statistical inference about a possible basic neighborhood Γ based on the blocks $a(\Gamma) \in A^\Gamma$ appearing in the sample $x(\Lambda_n)$. For technical reason, we will consider only such blocks whose center is in a subregion $\bar{\Lambda}_n$ of

Λ_n, consisting of those sites $i \in \Lambda_n$ for which the ball with center i and radius $\log^{\frac{1}{2d}} |\Lambda_n|$ also belongs to Λ_n:

$$\bar{\Lambda}_n = \left\{ i \in \Lambda_n : \left\{ j \in \mathbb{Z}^d : \|i - j\| \le \log^{\frac{1}{2d}} |\Lambda_n| \right\} \subseteq \Lambda_n \right\},$$

see Fig.3.1. Logarithms are to the base e. Our only assumptions about the sample regions Λ_n will be that

$$\Lambda_1 \subset \Lambda_2 \subset \ldots; \quad |\Lambda_n| \,/ \left|\bar{\Lambda}_n\right| \to 1.$$

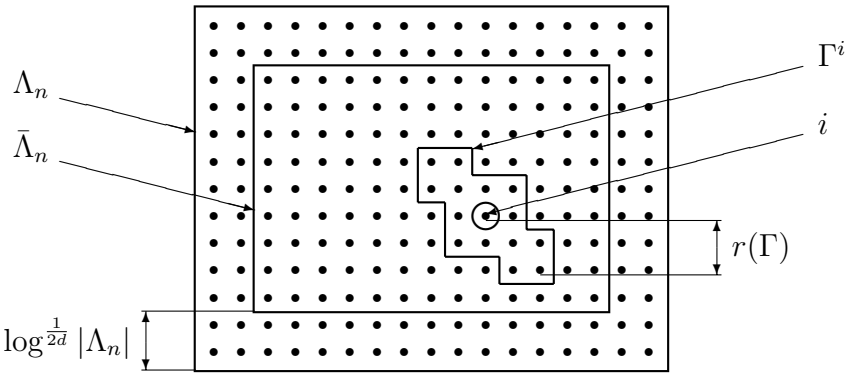

Figure 3.1: The Γ-neighborhood of the site i, and the sample region Λ_n.

For each block $a(\Gamma) \in A^\Gamma$, let $N_n(a(\Gamma))$ denote the number of occurrences of the block $a(\Gamma)$ in the sample $x(\Lambda_n)$ with the center in $\bar{\Lambda}_n$:

$$N_n(a(\Gamma)) = \left| \left\{ i \in \bar{\Lambda}_n : \Gamma^i \subseteq \Lambda_n,\, x(\Gamma^i) = a(\Gamma) \right\} \right|.$$

The blocks corresponding to Γ-neighborhoods completed with their centers, will be denoted briefly by $a(\Gamma, 0)$. Similarly as above, for each $a(\Gamma, 0) \in A^{\Gamma \cup \{0\}}$ we write

$$N_n(a(\Gamma, 0)) = \left| \left\{ i \in \bar{\Lambda}_n : \Gamma^i \subseteq \Lambda_n,\, x(\Gamma^i \cup \{i\}) = a(\Gamma, 0) \right\} \right|.$$

The notation $a(\Gamma, 0) \in x(\Lambda_n)$ will mean that $N_n(a(\Gamma, 0)) \ge 1$.

The restriction $\Gamma^i \subseteq \Lambda_n$ in the above definitions is automatically satisfied if $r(\Gamma) \le \log^{\frac{1}{2d}} |\Lambda_n|$. Hence, the same number of blocks is taken into account for all neighborhoods, except for very large ones:

$$\sum_{a(\Gamma) \in A^\Gamma} N_n(a(\Gamma)) = \left|\bar{\Lambda}_n\right|, \quad \text{if } r(\Gamma) \le \log^{\frac{1}{2d}} |\Lambda_n|.$$

For Markov random fields, the likelihood function cannot be explicitly determined. We shall use instead the pseudo-likelihood defined below.

Given the sample $x(\Lambda_n)$, the *pseudo-likelihood* function associated with a neighborhood Γ is the following function of a matrix Q'_Γ regarded as the one-point specification of a hypothetical Markov random field for which Γ is a Markov neighborhood:

(3.2)
$$\mathrm{PL}_\Gamma(x(\Lambda_n), Q'_\Gamma) = \prod_{i\in\bar{\Lambda}_n} Q'_\Gamma(x(i)|\, x(\Gamma^i)) = \prod_{a(\Gamma,0)\in x(\Lambda_n)} Q'_\Gamma(a(0)|\, a(\Gamma))^{N_n(a(\Gamma,0))}.$$

We note that not all matrices Q'_Γ satisfying

$$\sum_{a\in A} Q'_\Gamma(a(0)|\, a(\Gamma)) = 1, \quad a(\Gamma) \in A^\Gamma$$

are possible one-point specifications, the elements of a one-point specification matrix have to satisfy several algebraic relations not entered here. Still, we define the pseudo-likelihood also for Q'_Γ not satisfying those relations, even admitting some elements of Q'_Γ to be 0.

The maximum of this pseudo-likelihood is attained for $Q'_\Gamma(a(0)|\, a(\Gamma)) = \frac{N_n(a(\Gamma,0))}{N_n(a(\Gamma))}$. Thus, given the sample $x(\Lambda_n)$, the logarithm of the *maximum pseudo-likelihood* for the neighborhood Γ is

$$\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) = \sum_{a(\Gamma,0)\in x(\Lambda_n)} N_n(a(\Gamma,0)) \log \frac{N_n(a(\Gamma,0))}{N_n(a(\Gamma))}. \tag{3.3}$$

Now we are able to formalize a criterion to the analogy of the Bayesian Information Criterion that can be calculated from the sample.

Definition 3.2. *Given a sample $x(\Lambda_n)$, the Pseudo-Bayesian Information Criterion, shortly* PIC*, for the neighborhood Γ is*

$$\mathrm{PIC}_\Gamma(x(\Lambda_n)) = -\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) + |A|^{|\Gamma|} \log |\Lambda_n|\,.$$

Remark 3.3. *In our penalty term, the number $|A|^{|\Gamma|}$ of possible blocks $a(\Gamma) \in A^\Gamma$ replaces "half the number of free parameters" appearing in BIC, for which number no simple formula is available. Note that our results remain valid, with the same proofs, if the above penalty term is multiplied by any $c > 0$.* □

The PIC estimator of the basic neighborhood Γ_0 is defined as that hypothetical Γ for which the value of the criterion is minimal. An important feature of our estimator is that the family of hypothetical Γ's is allowed to extend as $n \to \infty$, thus no a priori upper bound for the size of the unknown Γ_0 is needed. Our main result says the PIC estimator is strongly consistent if the hypothetical Γ's are those with $r(\Gamma) \le r_n$, where r_n grows sufficiently slowly.

We mean by strong consistency that the estimated basic neighborhood equals Γ_0 eventually almost surely as $n \to \infty$. Here and in the sequel, "eventually almost surely" means that with probability 1 there exists a threshold n_0 (depending on the infinite realization $x(\mathbb{Z}^d)$) such that the claim holds for all $n \geq n_0$.

Theorem 3.4. *The PIC-estimator*

$$\widehat{\Gamma}_{\mathrm{PIC}}(x(\Lambda_n)) = \arg \min_{\Gamma:\, r(\Gamma) \leq r_n} \mathrm{PIC}_\Gamma(x(\Lambda_n)),$$

with

$$r_n = o\left(\log^{\frac{1}{2d}} |\Lambda_n|\right),$$

satisfies

$$\widehat{\Gamma}_{\mathrm{PIC}}(x(\Lambda_n)) = \Gamma_0,$$

eventually almost surely as $n \to \infty$.

Proof. Theorem 3.4 follows from Propositions 3.10 and 3.11 below. □

Remark 3.5. *Actually, the assertion will be proved for r_n equal to a constant times $\log^{\frac{1}{2d}} \left|\bar{\Lambda}_n\right|$. However, as this constant depends on the unknown distribution Q, the consistency can be guaranteed only when*

$$r_n = o\left(\log^{\frac{1}{2d}} \left|\bar{\Lambda}_n\right|\right) = o\left(\log^{\frac{1}{2d}} |\Lambda_n|\right).$$

It remains open whether consistency holds when the hypothetical neighborhoods are allowed to grow faster, or even without any condition on the hypothetical neighborhoods. □

As a consequence of the above, we are able to construct a strongly consistent estimator of the one-point specification Q_{Γ_0}.

Corollary 3.6. *The empirical estimator of the one-point specification,*

$$\widehat{Q}_{\widehat{\Gamma}}(a(0)|\, a(\widehat{\Gamma})) = \frac{N_n(a(\widehat{\Gamma}, 0))}{N_n(a(\widehat{\Gamma}))}\ , \quad a(0) \in A,\, a(\widehat{\Gamma}) \in A^{\widehat{\Gamma}},$$

converges to the true Q_{Γ_0} almost surely as $n \to \infty$, where $\widehat{\Gamma}$ is the PIC estimator $\widehat{\Gamma}_{\mathrm{PIC}}$.

Proof. Immediate from Theorem 3.4 and Proposition 3.7 below. □

3.3 The typicality result

Proposition 3.7. *Simultaneously for all Markov neighborhoods with* $r(\Gamma) \le \alpha^{\frac{1}{2d}} \log^{\frac{1}{2d}} \left|\bar{\Lambda}_n\right|$, *and blocks* $a(\Gamma, 0) \in A^{\Gamma \cup \{0\}}$,

$$\left| \frac{N_n(a(\Gamma, 0))}{N_n(a(\Gamma))} - Q(a(0)|\, a(\Gamma)) \right| < \sqrt{\frac{\kappa \log N_n(a(\Gamma))}{N_n(a(\Gamma))}},$$

eventually almost surely as $n \to \infty$, *if*

$$0 < \alpha \le 1, \quad \kappa > 2^{3d} e\, \alpha \log(|A|^2 + 1).$$

To prove this proposition we will use an idea similar to the "coding technique" of Besag (1974), namely we partition $\bar{\Lambda}_n$ into subsets $\bar{\Lambda}_n^k$ such that the random variables at the sites $i \in \bar{\Lambda}_n^k$ are conditionally independent given the values of those at the other sites. First, we introduce some further notation. Let

$$R_n = \left\lfloor \alpha^{\frac{1}{2d}} \left\lceil \log \left|\bar{\Lambda}_n\right| \right\rceil^{\frac{1}{2d}} \right\rfloor. \tag{3.4}$$

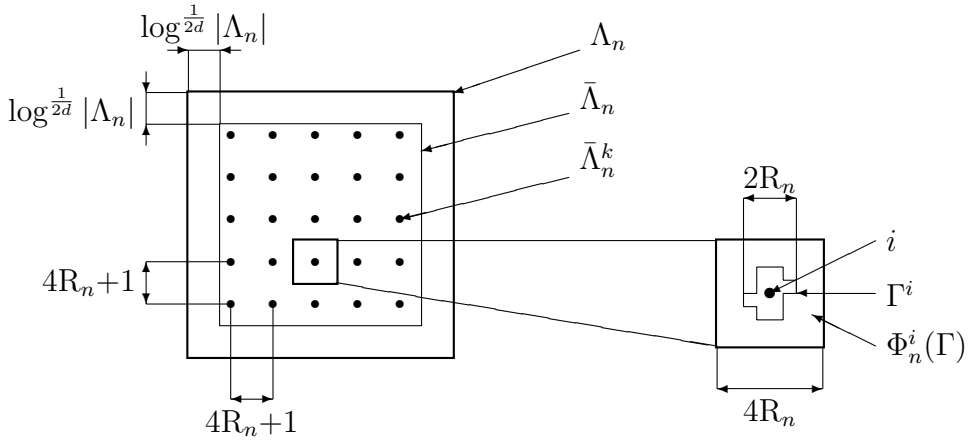

Figure 3.2: The sieve $\bar{\Lambda}_n^k$.

We partition the region $\bar{\Lambda}_n$ by intersecting it with sublattices of $\mathbb{Z}^d$ such that the distance between sites in a sublattice is $4R_n + 1$. The intersections of $\bar{\Lambda}_n$ with these sublattices will be called sieves. Indexed by the offset k relative to the origin 0, the sieves are

$$\bar{\Lambda}_n^k = \left\{\, i \in \bar{\Lambda}_n : i = k + (4R_n + 1)v,\, v \in \mathbb{Z}^d \,\right\}, \quad \|k\| \le 2R_n,$$

see Fig.3.2. For a neighborhood Γ, let $N_n^k(a(\Gamma))$ denote the number of occurrences of the block $a(\Gamma) \in A^\Gamma$ in the sample $x(\Lambda_n)$ with center in $\bar{\Lambda}_n^k$:

$$N_n^k(a(\Gamma)) = \left|\left\{\, i \in \bar{\Lambda}_n^k : \Gamma^i \subseteq \Lambda_n,\, x(\Gamma^i) = a(\Gamma) \,\right\}\right|.$$

Similarly, let

$$N_n^k(a(\Gamma,0)) = \left|\left\{\, i \in \bar{\Lambda}_n^k : \Gamma^i \subseteq \Lambda_n,\, x(\Gamma^i \cup \{i\}) = a(\Gamma,0) \,\right\}\right|.$$

Clearly,

$$N_n(a(\Gamma)) = \sum_{k:\,\|k\|\le 2R_n} N_n^k(a(\Gamma)) \quad \text{and} \quad N_n(a(\Gamma,0)) = \sum_{k:\,\|k\|\le 2R_n} N_n^k(a(\Gamma,0)).$$

The notation $a(\Gamma) \in x(\Lambda_n^k)$ will mean that $N_n^k(a(\Gamma)) \ge 1$.

Denote by $\Phi_n(\Gamma)$ the set of sites outside the neighborhood Γ whose norm is at most $2R_n$:

$$\Phi_n(\Gamma) = \left\{\, i \in \mathbb{Z}^d : \|i\| \le 2R_n,\, i \notin \Gamma \,\right\},$$

see Fig.3.2. $\Phi_n^i(\Gamma)$ denotes the translate of $\Phi_n(\Gamma)$ when 0 is translated to i.

For a finite region $\Xi \subset \mathbb{Z}^d$, conditional probabilities on the condition $X(\Xi) = x(\Xi) \in A^\Xi$, will be denoted briefly by $\mathrm{Prob}\{\,\cdot \mid x(\Xi)\,\}$.

In the following Lemma the neighborhoods Γ need not be Markov neighborhoods.

Lemma 3.8. *Simultanously for all sieves k, neighborhoods Γ with $r(\Gamma) \le R_n$, and blocks $a(\Gamma) \in A^\Gamma$,*

$$(1+\varepsilon)\log N_n^k(a(\Gamma)) \ge \log\left|\bar{\Lambda}_n\right|,$$

eventually almost surely as $n \to \infty$, where $\varepsilon > 0$ is an arbitrary constant.

Proof. As a consequence of Lemma 3.1, for any fixed sieve k and neighborhood Γ with $r(\Gamma) \le R_n$, the random variables $X(\Gamma^i)$, $i \in \bar{\Lambda}_n^k$ are conditionally independent given the values of the random variables in the rest of the sites of sample region Λ_n. By Lemma 3.20 in the Appendix,

$$Q\left(\, a(\Gamma) \mid a(\Phi_n(\Gamma))\,\right) \ge q_{\min}^{|\Gamma|}, \quad a(\Phi_n(\Gamma)) \in A^{\Phi_n(\Gamma)},$$

hence we can use the large deviation theorem of Lemma 3.18 in the Appendix with $p_* = q_{\min}^{|\Gamma|}$ to obtain

$$\mathrm{Prob}\left\{ \frac{N_n^k(a(\Gamma))}{\left|\bar{\Lambda}_n^k\right|} < \frac{1}{2}\, q_{\min}^{|\Gamma|} \;\middle|\; x\left(\Lambda_n \backslash \bigcup\nolimits_{i\in\bar{\Lambda}_n^k} \Gamma^i\right)\right\} \le \exp\left[-\left|\bar{\Lambda}_n^k\right| \frac{q_{\min}^{|\Gamma|}}{16} \right].$$

Hence also for the unconditional probabilities,

$$\mathrm{Prob}\left\{ \frac{N_n^k(a(\Gamma))}{\left|\bar{\Lambda}_n^k\right|} < \frac{1}{2}\, q_{\min}^{|\Gamma|} \right\} \le \exp\left[-\left|\bar{\Lambda}_n^k\right| \frac{q_{\min}^{|\Gamma|}}{16} \right].$$

Note that for $n \geq n_0$ (not depending on k) we have

$$\left|\bar{\Lambda}_n^k\right| \geq \frac{1}{2}\,\frac{\left|\bar{\Lambda}_n\right|}{(4R_n+1)^d} > \frac{\left|\bar{\Lambda}_n\right|}{(5R_n)^d}.$$

Using this and the consequence $|\Gamma| \leq (2R_n+1)^d < (3R_n)^d$ of $r(\Gamma) \leq R_n$, the last probability bound implies for $n \geq n_0$

$$\text{Prob}\left\{\frac{N_n^k(a(\Gamma))}{\left|\bar{\Lambda}_n\right|} < \frac{q_{\min}^{(3R_n)^d}}{2\,(5R_n)^d}\right\} \leq \exp\left[-\left|\bar{\Lambda}_n\right|\frac{q_{\min}^{(3R_n)^d}}{16\,(5R_n)^d}\right].$$

Using the union bound and Lemma 3.21 in the Appendix, it follows that

$$\begin{aligned}\text{Prob}\Bigg\{&\frac{N_n^k(a(\Gamma))}{\left|\bar{\Lambda}_n\right|} < \frac{q_{\min}^{(3R_n)^d}}{2\,(5R_n)^d}\,,\\ &\qquad\text{for some } k,\Gamma,a(\Gamma) \text{ with } \|k\| \leq 2R_n,\ r(\Gamma) \leq R_n,\ a(\Gamma) \in A^\Gamma\Bigg\}\\ &\leq \exp\left[-\left|\bar{\Lambda}_n\right|\frac{q_{\min}^{(3R_n)^d}}{16\,(5R_n)^d}\right]\cdot(4R_n+1)^d\cdot(|A|^2+1)^{(2R_n+1)^d/2}.\end{aligned}$$

Recalling (3.4), this is summable in n, thus the Borel – Cantelli lemma gives

$$N_n^k(a(\Gamma)) \geq \left|\bar{\Lambda}_n\right|\frac{q_{\min}^{3^d\alpha^{1/2}\left(1+\log\left|\bar{\Lambda}_n\right|\right)^{1/2}}}{2\cdot 5^d\alpha^{1/2}\left(1+\log\left|\bar{\Lambda}_n\right|\right)^{1/2}},$$

eventually almost surely as $n \to \infty$, simultanously for all sieves k, neighborhoods Γ with $r(\Gamma) \leq R_n$, and blocks $a(\Gamma) \in A^\Gamma$. This proves the Lemma. □

Lemma 3.9. *Simultaneously for all sieves k, Markov neighborhoods Γ with $r(\Gamma) \leq R_n$, and blocks $a(\Gamma,0) \in A^{\Gamma\cup\{0\}}$,*

$$\left|\frac{N_n^k(a(\Gamma,0))}{N_n^k(a(\Gamma))} - Q(a(0)|\,a(\Gamma))\right| < \sqrt{\frac{\delta\log^{\frac{1}{2}}N_n^k(a(\Gamma))}{N_n^k(a(\Gamma))}}\,,$$

eventually almost surely as $n \to \infty$, if

$$\delta > 2^d e\,\alpha^{\frac{1}{2}}\log(|A|^2+1).$$

Proof. Given a sieve k, a Markov neighborhood Γ, and a block $a(\Gamma,0)$, the difference $N_n^k(a(\Gamma,0)) - N_n^k(a(\Gamma))\,Q(a(0)|\,a(\Gamma))$ equals

$$Y_n = \sum_{i\in\bar{\Lambda}_n^k:\,x(\Gamma^i)=a(\Gamma)} [\,\mathbb{I}(X(i)=a(0)) - Q(a(0)|\,a(\Gamma))\,]\,,$$

where $\mathbb{I}(\,\cdot\,)$ denotes indicator function, hence the claimed inequality is equivalent to

$$-\sqrt{N_n^k(a(\Gamma))\delta \log^{\frac{1}{2}} N_n^k(a(\Gamma))} < Y_n < \sqrt{N_n^k(a(\Gamma))\delta \log^{\frac{1}{2}} N_n^k(a(\Gamma))}.$$

We will prove that the last inequalities hold eventually almost surely as $n \to \infty$, simultaneously for all sieves k, Markov neighborhoods Γ with $r(\Gamma) \le R_n$, and blocks $a(\Gamma, 0) \in A^{\Gamma \cup \{0\}}$. We concentrate on the second inequality, the proof for the first one is similar.

Denote

$$G_j(k, a(\Gamma, 0)) = \left\{ \max_{n \in \mathcal{N}_j(k, a(\Gamma))} Y_n \ge \sqrt{e^j \delta j^{\frac{1}{2}}} \right\},$$

where

$$\mathcal{N}_j(k, a(\Gamma)) = \left\{ n : \, e^j < N_n^k(a(\Gamma)) \le e^{j+1}, \; (1+\varepsilon) \log N_n^k(a(\Gamma)) \ge \log \left|\bar{\Lambda}_n\right| \right\};$$

if $n \in \mathcal{N}_j(k, a(\Gamma))$ then, by (3.4),

$$R_n = \left\lfloor \alpha^{\frac{1}{2d}} \left\lceil \log \left|\bar{\Lambda}_n\right| \right\rceil^{\frac{1}{2d}} \right\rfloor \le \alpha^{\frac{1}{2d}} \left(1 + (1+\varepsilon)(j+1)\right)^{\frac{1}{2d}} \stackrel{\text{def}}{=} R^{(j)}. \tag{3.5}$$

The claimed inequality $Y_n < \sqrt{N_n^k(a(\Gamma))\delta \log^{\frac{1}{2}} N_n^k(a(\Gamma))}$ holds for each n with $e^j < N_n^k(a(\Gamma)) \le e^{j+1}$ if

$$\max_{n:\, e^j < N_n^k(a(\Gamma)) \le e^{j+1}} Y_n < \sqrt{e^j \delta j^{\frac{1}{2}}}.$$

By Lemma 3.8, the condition $(1+\varepsilon) \log N_n^k(a(\Gamma)) \ge \log \left|\bar{\Lambda}_n\right|$ in the definition of $\mathcal{N}_j(k, a(\Gamma))$ is satisfied eventually almost surely, simultanously for all sieves k, neighborhoods Γ with $r(\Gamma) \le R_n$, and blocks $a(\Gamma) \in A^{\Gamma}$. Hence it suffices to prove that the following holds with probability one: the union of the events $G_j(k, a(\Gamma, 0))$ for all k with $\|k\| \le 2R^{(j)}$, all $\Gamma \supseteq \Gamma_0$ with $r(\Gamma) \le R^{(j)}$, and all $a(\Gamma, 0) \in A^{\Gamma \cup \{0\}}$, obtains only for finitely many j.

As $n \in \mathcal{N}_j(k, a(\Gamma))$ implies $j < \log \left|\bar{\Lambda}_n\right| \le (1+\varepsilon)(j+1)$,

$$G_j(k, a(\Gamma, 0)) \subseteq \bigcup_{l=j}^{\lfloor (1+\varepsilon)(j+1) \rfloor} \left\{ \max_{n \in \mathcal{N}_{j,l}(k, a(\Gamma))} Y_n \ge \sqrt{e^j \delta j^{\frac{1}{2}}} \right\}, \tag{3.6}$$

where

$$\mathcal{N}_{j,l}(k, a(\Gamma)) = \left\{ n : \, e^j < N_n^k(a(\Gamma)) \le e^{j+1}, \; l < \log \left|\bar{\Lambda}_n\right| \le l+1 \right\}.$$

The random variables $X(i)$, $i \in \bar{\Lambda}_n^k$ are conditionally independent given the values of the random variables in their Γ-neighborhoods. Moreover, those $X(i)$'s

for which the same block $a(\Gamma)$ appears in their Γ-neighborhood, are also conditionally i.i.d.. Hence Y_n is the sum of $N_n^k(a(\Gamma))$ conditionally i.i.d. random variables with mean 0 and variance

$$\frac{1}{4} \geq D^2 = Q(a(0)|\, a(\Gamma))\,[\,1 - Q(a(0)|\, a(\Gamma))\,] \geq \frac{1}{2}\, q_{\min}.$$

As R_n is constant for n with $l < \log\left|\bar{\Lambda}_n\right| \leq l+1$, the corresponding Y_n's are actually partial sums of a sequence of $N_{n^*}^k(a(\Gamma)) \leq e^{j+1}$ such conditionally i.i.d. random variables, where n^* is the largest element of $\mathcal{N}_{j,l}(k, a(\Gamma))$. Therefore, using Lemma 3.19 in the Appendix with $\mu = \mu_j = (1-\eta)\sqrt{e^{-1}\delta j^{1/2}}$, where $\eta > 0$ is an arbitrary constant, we have

$$\begin{aligned}
&\text{Prob}\left\{ \max_{n\,\in\mathcal{N}_{j,l}(k,a(\Gamma))} Y_n \geq \sqrt{e^j \delta j^{\frac{1}{2}}} \;\middle|\; x\left(\bigcup_{i\in\bar{\Lambda}_n^k:\, x(\Gamma^i)=a(\Gamma)} \Gamma^i\right)\right\} \\
&\quad\leq \text{Prob}\left\{ \max_{n\,\in\mathcal{N}_{j,l}(k,a(\Gamma))} Y_n \geq D\sqrt{e^{j+1}}\left((1-\eta)\sqrt{e^{-1}\delta j^{\frac{1}{2}}} + 2\right)\right. \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\left. \;\middle|\; x\left(\bigcup_{i\in\bar{\Lambda}_n^k:\, x(\Gamma^i)=a(\Gamma)} \Gamma^i\right)\right\} \\
&\quad\leq \frac{8}{3}\,\exp\left[-\frac{\mu_j^2}{2\left(1+\frac{\mu_j}{2D\sqrt{e^{j+1}}}\right)^2}\right]
\end{aligned}$$

On account of $\lim\limits_{j\to\infty} \mu_j/(2D\sqrt{e^{j+1}}) = 0$, the last bound can be continued, for $j > j_0$, as

$$\leq \frac{8}{3}\,\exp\left[-\frac{(1-\eta)^2}{2e(1+\eta)}\,\delta j^{\frac{1}{2}}\right].$$

This bound also holds for the unconditional probabilities, hence we obtain from (3.6)

$$\begin{aligned}
\text{Prob}\,\{\, G_j(k, a(\Gamma, 0))\,\} &\leq (\varepsilon j + 2)\cdot\frac{8}{3}\,\exp\left[-\frac{(1-\eta)^2}{2e(1+\eta)}\,\delta j^{\frac{1}{2}}\right] \\
&\leq \exp\left[-\frac{(1-\eta)^3}{2e(1+\eta)}\,\delta j^{\frac{1}{2}}\right].
\end{aligned}$$

To bound the number of all admissible k, Γ, $a(\Gamma, 0)$ (recall the conditions $\|k\| \leq 2R^{(j)}$, $r(\Gamma) \leq R^{(j)}$, with $R^{(j)}$ defined in (3.5)), note that the number of possible k's is bounded by

$$(4R^{(j)} + 1)^d \leq (4+\rho)^d \alpha^{\frac{1}{2}} (1+\varepsilon)^{\frac{1}{2}} (j+1)^{\frac{1}{2}},$$

and, by Lemma 3.21 in the Appendix, the number of possible blocks $a(\Gamma, 0)$ with $r(\Gamma) \le R^{(j)}$ is bounded by

$$(|A|^2+1)^{(2R^{(j)}+1)^d/2} < (|A|^2+1)^{(1+\rho)^d 2^{d-1}\alpha^{1/2}(1+\varepsilon)^{1/2}(j+1)^{1/2}}.$$

Combining the above bounds, we get for the probability of the union of the events $G_j(k, a(\Gamma,0))$ for all admissible k, Γ, $a(\Gamma,0)$ the bound

$$\exp\Bigg[-\frac{(1-\eta)^3}{2e(1+\eta)}\,\delta j^{\frac12} + [\log(|A|^2+1)](1+\rho)^d 2^{d-1}\alpha^{\frac12}(1+\varepsilon)^{\frac12}(j+1)^{\frac12} + O\left(\log j^{\frac12}\right)\Bigg].$$

This is summable in j, if we choose η, ε, ρ sufficiently small, and $\delta/(2e) > 2^{d-1}\alpha^{\frac12}\log(|A|^2+1)$, that is, if $\delta > 2^d e\,\alpha^{\frac12}\log(|A|^2+1)$. □

Proof of Proposition 3.7. Using Lemma 3.9,

$$\begin{aligned}
&\left|\frac{N_n(a(\Gamma,0))}{N_n(a(\Gamma))} - Q(a(0)|\,a(\Gamma))\right| \\
&\le \sum_{k:\,\|k\|\le 2R_n}\left|\frac{N_n^k(a(\Gamma,0))}{N_n^k(a(\Gamma))} - Q(a(0)|\,a(\Gamma))\right|\cdot\frac{N_n^k(a(\Gamma))}{N_n(a(\Gamma))} \\
&< \sum_{k:\,\|k\|\le 2R_n}\sqrt{\frac{\delta\log^{1/2}N_n^k(a(\Gamma))}{N_n^k(a(\Gamma))}}\cdot\frac{N_n^k(a(\Gamma))}{N_n(a(\Gamma))}
\end{aligned}$$

eventually almost surely as $n\to\infty$. By Jensen's inequality and $N_n^k(a(\Gamma)) \le N_n(a(\Gamma))$, this can be continued as

$$\le \sqrt{\frac{\delta\,(4R_n+1)^d\,\log^{1/2}N_n(a(\Gamma))}{N_n(a(\Gamma))}}.$$

Since by (3.4) and Lemma 3.8, we have for any ε, $\rho > 0$ and n sufficiently large

$$(4R_n+1)^d \le \left(4\alpha^{\frac{1}{2d}}\left(1+\log\left|\bar\Lambda_n\right|\right)^{\frac{1}{2d}}+1\right)^d \le (4+\rho)^d\alpha^{1/2}(1+\varepsilon)^{1/2}\log^{1/2}N_n(a(\Gamma)),$$

eventually almost surely as $n\to\infty$. This completes the proof. □

3.4 The overestimation

Proposition 3.10. *Eventually almost surely as $n\to\infty$,*

$$\widehat\Gamma_{\mathrm{PIC}}(x(\Lambda_n)) \notin \{\,\Gamma : \Gamma \supset \Gamma_0\,\},$$

whenever r_n in Theorem 3.4 is equal to R_n in (3.4) with

$$\alpha < \frac{q_{min}}{2^{3d}e}\,\frac{|A|-1}{|A|^2\log(|A|^2+1)}.$$

Proof. We have to prove that simultaneously for all neighborhoods $\Gamma \supset \Gamma_0$ with $r(\Gamma) \le R_n$,

$$\mathrm{PIC}_\Gamma(x(\Lambda_n)) - \mathrm{PIC}_{\Gamma_0}(x(\Lambda_n)) > 0, \tag{3.7}$$

eventually almost surely as $n \to \infty$.

The left hand side

$$-\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) + |A|^{|\Gamma|} \log|\Lambda_n| + \log \mathrm{MPL}_{\Gamma_0}(x(\Lambda_n)) - |A|^{|\Gamma_0|} \log|\Lambda_n|$$

is bounded below by

$$-\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) + \log \mathrm{PL}_{\Gamma_0}(x(\Lambda_n), Q_{\Gamma_0}) + \left(1 - \frac{1}{|A|}\right) |A|^{|\Gamma|} \log|\Lambda_n| \,.$$

Hence, it suffices to show that simultaneously for all neighborhoods $\Gamma \supset \Gamma_0$ with $r(\Gamma) \le R_n$,

$$\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) - \log \mathrm{PL}_{\Gamma_0}(x(\Lambda_n), Q_{\Gamma_0}) < \frac{|A|-1}{|A|} |A|^{|\Gamma|} \log|\Lambda_n| \,, \tag{3.8}$$

eventually almost surely as $n \to \infty$.

Now, for $\Gamma \supset \Gamma_0$ we have $\mathrm{PL}_{\Gamma_0}(x(\Lambda_n), Q_{\Gamma_0}) = \mathrm{PL}_\Gamma(x(\Lambda_n), Q_\Gamma)$, by the definition (3.2) of pseudo-likelihood, since Γ_0 is a Markov neighborhood. Thus, the left hand side of (3.8) equals

$$\begin{aligned}
&\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) - \log \mathrm{PL}_\Gamma(x(\Lambda_n), Q_\Gamma) \\
&= \sum_{a(\Gamma,0) \in x(\Lambda_n)} N_n(a(\Gamma,0)) \log \frac{N_n(a(\Gamma,0)) \,/\, N_n(a(\Gamma))}{Q(a(0)|\, a(\Gamma))} \\
&= \sum_{a(\Gamma) \in x(\Lambda_n)} N_n(a(\Gamma)) \sum_{a(0):\, a(\Gamma,0) \in x(\Lambda_n)} \frac{N_n(a(\Gamma,0))}{N_n(a(\Gamma))} \log \frac{N_n(a(\Gamma,0)) \,/\, N_n(a(\Gamma))}{Q(a(0)|\, a(\Gamma))}.
\end{aligned}$$

To bound the last expression, we use Proposition 3.7 and Lemma 3.22 in the Appendix, the latter applied with $P(a(0)) = \frac{N_n(a(\Gamma,0))}{N_n(a(\Gamma))}$, $Q(a(0)) = Q(a(0)|\, a(\Gamma))$. Thus we obtain the upper bound

$$\begin{aligned}
&\sum_{a(\Gamma) \in x(\Lambda_n)} N_n(a(\Gamma)) \frac{1}{q_{\min}} \sum_{a(0):\, a(\Gamma,0) \in x(\Lambda_n)} \left[\frac{N_n(a(\Gamma,0))}{N_n(a(\Gamma))} - Q(a(0)|\, a(\Gamma)) \right]^2 \\
&< \sum_{a(\Gamma) \in x(\Lambda_n)} N_n(a(\Gamma)) \frac{1}{q_{\min}} |A| \frac{\kappa \log N_n(a(\Gamma))}{N_n(a(\Gamma))} \le \frac{\kappa |A|}{q_{\min}} |A|^{|\Gamma|} \log\left|\bar{\Lambda}_n\right| ,
\end{aligned}$$

eventually almost surely as $n \to \infty$, simultaneously for all neighborhoods $\Gamma \supset \Gamma_0$ with $r(\Gamma) \le R_n$.

Hence, since $|\Lambda_n| \,/\left|\bar{\Lambda}_n\right| \to 1$, the assertion (3.8) holds whenever

$$\frac{\kappa |A|}{q_{\min}} < \frac{|A|-1}{|A|} \,,$$

which is equivalent to the bound on α in Proposition 3.10. □

3.5 The underestimation

Proposition 3.11. *Eventually almost surely as* $n \to \infty$,

$$\widehat{\Gamma}_{\mathrm{PIC}}(x(\Lambda_n)) \in \{\, \Gamma : \Gamma \supseteq \Gamma_0 \,\},$$

if r_n *in Theorem 3.4 is chosen as in Proposition 3.10.*

Proposition 3.11 will be proved using the lemmas below. Let us denote

$$\Psi_0 = \Big(\bigcup\nolimits_{i \in \Gamma_0} \Gamma_0^i \Big) \Big\backslash \Big(\Gamma_0 \cup \{0\} \Big).$$

Lemma 3.12. *The assertion of Proposition 3.7 holds also with* Γ *replaced by* $\Gamma \cup \Psi_0$, *where* Γ *is any (not necessarily Markov) neighborhood.*

Proof. As Proposition 3.7 was a consequence of Lemma 3.9, we have to check that the proof of that Lemma works when the Markov neighborhood Γ is replaced by $\Gamma \cup \Psi_0$, where Γ is any neighborhood. To this end, it suffices to show that conditioned on the values of all random variables in the $(\Gamma \cup \Psi_0)$-neighborhoods of the sites $i \in \bar{\Lambda}_n^k$, those $X(i)$, $i \in \bar{\Lambda}_n^k$ are conditionally i.i.d. for which the same block $a(\Gamma \cup \Psi_0)$ appears in the $(\Gamma \cup \Psi_0)$-neighborhood of i. This follows from Lemma 3.16 in the Appendix, with $\Delta = \Gamma_0 \cup \{0\}$ and $\Psi = \Psi_0$. □

Lemma 3.13. *Simultanously for all neighborhoods* $\Gamma \not\supseteq \Gamma_0$ *with* $r(\Gamma) \le R_n$,

$$\mathrm{PIC}_{\Gamma \cup \Psi_0}(x(\Lambda_n)) > \mathrm{PIC}_{(\Gamma \cap \Gamma_0) \cup \Psi_0}(x(\Lambda_n)),$$

eventually almost surely as $n \to \infty$.

Proof. The claimed inequality is analogous to (3.7) in the proof of Proposition 3.10, the roles of $\Gamma \supset \Gamma_0$ there played by $\Gamma \cup \Psi_0 \supset (\Gamma \cap \Gamma_0) \cup \Psi_0$. Its proof is the same as that of (3.7), using Lemma 3.12 instead of Proposition 3.7. Indeed, the basic neighborhood property of Γ_0 was used in that proof only to show that $\mathrm{PL}_{\Gamma_0}(x(\Lambda_n), Q_{\Gamma_0}) = \mathrm{PL}_{\Gamma}(x(\Lambda_n), Q_{\Gamma})$. The analogue of this identity, namely

$$\mathrm{PL}_{(\Gamma \cap \Gamma_0) \cup \Psi_0}(x(\Lambda_n), Q_{(\Gamma \cap \Gamma_0) \cup \Psi_0}) = \mathrm{PL}_{\Gamma \cup \Psi_0}(x(\Lambda_n), Q_{\Gamma \cup \Psi_0}),$$

follows from Lemma 3.16 in the Appendix, with $\Delta = \Gamma_0 \cup \{0\}$ and $\Psi = \Psi_0$. □

For the next lemma, we introduce some further notation.

The set of all probability distributions on $A^{\mathbb{Z}^d}$, equipped with the weak topology, is a compact Polish space; let d denote a metric that metrizes it. Let $\mathcal{Q}^G$ denote the (compact) set of Gibbs distributions with the one-point specification Q_{Γ_0}.

For a sample $x(\Lambda_n)$, define the empirical distribution on $A^{\mathbb{Z}^d}$ by

$$R_{x,n} = \frac{1}{\left|\bar{\Lambda}_n\right|} \sum_{i \in \bar{\Lambda}_n} \delta_{x_n^i},$$

where $x_n \in A^{\mathbb{Z}^d}$ is the extension of the sample $x(\Lambda_n)$ to the whole lattice with $x_n(j)$ equal to a constant $a \in A$ for $j \in \mathbb{Z}^d \backslash \Lambda_n$, x_n^i denotes the translate of x_n when 0 is translated to i, and δ_x is the Dirac mass at $x \in A^{\mathbb{Z}^d}$.

Lemma 3.14. *With probability 1,* $d\left(R_{x,n}, \mathcal{Q}^G\right) \to 0$.

Proof. Fix a realization $x(\mathbb{Z}^d)$ for which Proposition 3.7 holds.

It suffices to show that for any subsequence n_k such that R_{x,n_k} converges, its limit $R_{x,0}$ belongs to $\mathcal{Q}^G$.

Let Γ' be any neighborhood. For n sufficiently large, the $(\Gamma' \cup \{0\})$-marginal of $R_{x,n}$ is equal to

$$\left\{ \frac{N_n(a(\Gamma', 0))}{\left|\bar{\Lambda}_n\right|} \ , \ a(\Gamma', 0) \in A^{\Gamma' \cup \{0\})} \right\},$$

hence $R_{x,n_k} \to R_{x,0}$ implies

$$\frac{N_{n_k}(a(\Gamma', 0))}{\left|\bar{\Lambda}_{n_k}\right|} \longrightarrow R_{x,0}(a(\Gamma', 0)) \tag{3.9}$$

for all $a(\Gamma', 0) \in A^{\Gamma' \cup \{0\}}$. This and summation for $a(0) \in A$, imply

$$\frac{N_{n_k}(a(\Gamma', 0))}{N_{n_k}(a(\Gamma'))} \longrightarrow R_{x,0}(a(0) | \, a(\Gamma')).$$

As Proposition 3.7 holds for the realization $x(\mathbb{Z}^d)$, it follows that if Γ' is a Markov neighborhood then

$$R_{x,0}(a(0) | \, a(\Gamma')) = Q(a(0) | \, a(\Gamma')) = Q_{\Gamma_0}(a(0) | \, a(\Gamma_0)).$$

For any finite region $\Delta \supset \Gamma_0$ with $0 \notin \Delta$, the last equation for a neighborhood $\Gamma' \supset \Delta$ implies that

$$R_{x,0}(a(0) | \, a(\Delta)) = Q_{\Gamma_0}(a(0) | \, a(\Gamma_0)) \quad \text{if } \Delta \supset \Gamma_0,\ 0 \notin \Delta.$$

To prove $R_{x,0} \in \mathcal{Q}^G$ it remains to show that, in addition, $R_{x,0}(a(i) | \, a(\Delta^i)) = Q_{\Gamma_0}(a(i) | \, a(\Gamma_0^i))$. Actually, we show that $R_{x,0}$ is translation invariant. Indeed, given a finite region $\Delta \subset \mathbb{Z}^d$ and its translate Δ^i, take a neighborhood Γ' with $\Delta \cup \Delta^i \subseteq \Gamma' \cup \{0\}$, and consider the sum of the counts $N_n(a(\Gamma', 0))$ for all blocks $a(\Gamma', 0) = \{\, a(j) : j \in \Gamma' \cup \{0\} \,\}$ with $\{\, a(j) : j \in \Delta \,\}$ equal to a fixed $|\Delta|$-tuple and the similar sum with $\{\, a(j) : j \in \Delta^i \,\}$ equal to the same $|\Delta|$-tuple. If $\|i\| < \log^{1/(2d)} \left|\bar{\Lambda}_n\right|$, the difference of these sums is at most $|\Lambda_n| - \left|\bar{\Lambda}_n\right|$, hence the translation invariance of $R_{x,0}$ follows by (3.9). $\square$

Lemma 3.15. *Uniformly for all neighborhoods* Γ *not containing* Γ_0*,*

$$-\log \mathrm{MPL}_{(\Gamma\cap\Gamma_0)\cup\Psi_0}(x(\Lambda_n)) > -\log \mathrm{MPL}_{\Gamma_0}(x(\Lambda_n)) + c\left|\bar{\Lambda}_n\right|,$$

eventually almost surely as $n \to \infty$*, where* $c > 0$ *is a constant.*

Proof. Given a realization $x \in A^{\mathbb{Z}^d}$ with the property in Lemma 3.14, there exists a sequence $Q_{x,n}$ in $\mathcal{Q}^G$ with

$$d\left(R_{x,n}, Q_{x,n}\right) \to 0,$$

and consequently

$$\frac{N_n(a(\Delta))}{\left|\bar{\Lambda}_n\right|} - Q_{x,n}(a(\Delta)) \longrightarrow 0 \tag{3.10}$$

for each finite region $\Delta \subset \mathbb{Z}^d$ and $a(\Delta) \in A^{\Delta}$.

Next, let Γ be a neighborhood with $\Gamma \not\supseteq \Gamma_0$. By (3.3),

$$\begin{aligned}
&-\frac{1}{\left|\bar{\Lambda}_n\right|}\log \mathrm{MPL}_{(\Gamma\cap\Gamma_0)\cup\Psi_0}(x(\Lambda_n)) \\
&= -\frac{1}{\left|\bar{\Lambda}_n\right|}\sum_{a((\Gamma\cap\Gamma_0)\cup\Psi_0,0)\in x(\Lambda_n)} N_n(a((\Gamma\cap\Gamma_0)\cup\Psi_0,0))\log\frac{N_n(a((\Gamma\cap\Gamma_0)\cup\Psi_0,0))}{N_n(a((\Gamma\cap\Gamma_0)\cup\Psi_0))}.
\end{aligned}$$

Applying (3.10) to $\Delta = (\Gamma\cap\Gamma_0)\cup\Psi_0\cup\{0\}$, it follows that the last expression is arbitrary close to

$$\begin{aligned}
&-\sum_{a((\Gamma\cap\Gamma_0)\cup\Psi_0\cup\{0\})} Q_{x,n}(a((\Gamma\cap\Gamma_0)\cup\Psi_0,0))\log Q_{x,n}(a(0)|\, a((\Gamma\cap\Gamma_0)\cup\Psi_0)) \\
&= H_{Q_{x,n}}(X(0)|\, X((\Gamma\cap\Gamma_0)\cup\Psi_0)),
\end{aligned}$$

if n is sufficiently large, where $H_{Q_{x,n}}(\,\cdot\,|\,\cdot\,)$ denotes conditional entropy, when the underlying distribution is $Q_{x,n}$. Similarly, $-\left(1/\left|\bar{\Lambda}_n\right|\right)\log \mathrm{MPL}_{\Gamma_0}(x(\Lambda_n))$ is arbitrary close to $H_{Q_{x,n}}(X(0)|\, X(\Gamma_0))$, that equals $H_{Q_{x,n}}(X(0)|\, X(\Gamma_0\cup\Psi_0))$ since Γ_0 is a Markov neighborhood.

It is known that $H_{Q'}(X(0)|\, X((\Gamma\cap\Gamma_0)\cup\Psi_0)) \geq H_{Q'}(X(0)|\, X(\Gamma_0\cup\Psi_0))$ for any distribution Q'. The proof of the Lemma will be complete if we show that, in addition, there exists a constant $\xi > 0$ (depending on $\Gamma\cap\Gamma_0$) such that for every Gibbs distribution $Q^G \in \mathcal{Q}^G$

$$H_{Q^G}(X(0)|\, X((\Gamma\cap\Gamma_0)\cup\Psi_0)) - H_{Q^G}(X(0)|\, X(\Gamma_0\cup\Psi_0)) > \xi.$$

The indirect assumption that the left hand side goes to 0 for some sequence of Gibbs distributions in $\mathcal{Q}^G$ implies, using the compactness of $\mathcal{Q}^G$, that

$$H_{Q_0^G}(X(0)|\, X((\Gamma\cap\Gamma_0)\cup\Psi_0)) = H_{Q_0^G}(X(0)|\, X(\Gamma_0\cup\Psi_0)),$$

for the limit $Q_0^G \in \mathcal{Q}^G$ of a convergent subsequence. This equality implies

$$Q_0^G(a(0)|\, a((\Gamma \cap \Gamma_0) \cup \Psi_0)) = Q_0^G(a(0)|\, a(\Gamma_0 \cup \Psi_0))$$

for all $a(0) \in A$, $a(\Gamma_0 \cup \Psi_0) \in A^{\Gamma_0 \cup \Psi_0}$. By Lemma 3.16 in the Appendix, these conditional probabilities are uniquely determined by the one-point specification Q_{Γ_0}, and the last equality implies

$$Q(a(i)|\, a((\Gamma \cap \Gamma_0)^i \cup \Psi_0^i)) = Q(a(i)|\, a(\Gamma_0^i \cup \Psi_0^i)) = Q_{\Gamma_0}(a(i)|\, a(\Gamma_0^i)).$$

According to Lemma 3.17 in the Appendix, this would imply $(\Gamma \cap \Gamma_0) \cup \Psi_0$ is a Markov neighborhood too, which is a contradiction, as $(\Gamma \cap \Gamma_0) \cup \Psi_0 \not\supseteq \Gamma_0$.

This completes the proof of the Lemma, because there is only a finite number of possible intersections $\Gamma \cap \Gamma_0$. □

Proof of Proposition 3.11. We have to show that

$$\mathrm{PIC}_\Gamma(x(\Lambda_n)) > \mathrm{PIC}_{\Gamma_0}(x(\Lambda_n)), \tag{3.11}$$

eventually almost surely as $n \to \infty$, for all neighborhoods Γ with $r(\Gamma) \le R_n$ that do not contain Γ_0.

Note that $\Gamma_1 \supseteq \Gamma_2$ implies $\mathrm{MPL}_{\Gamma_1}(x(\Lambda_n)) \ge \mathrm{MPL}_{\Gamma_2}(x(\Lambda_n))$, since $\mathrm{MPL}_\Gamma(x(\Lambda_n))$ is the maximizer in Q'_Γ of $\mathrm{PL}_\Gamma(x(\Lambda_n), Q'_\Gamma)$, see (3.2). Hence

$$-\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) \ge -\log \mathrm{MPL}_{\Gamma \cup \Psi_0}(x(\Lambda_n)),$$

for any neighborhood Γ.

Thus

$$\begin{aligned}\mathrm{PIC}_\Gamma(x(\Lambda_n)) &= -\log \mathrm{MPL}_\Gamma(x(\Lambda_n)) + |A|^{|\Gamma|} \log|\Lambda_n| \\ &\ge \mathrm{PIC}_{\Gamma \cup \Psi_0}(x(\Lambda_n)) - \left(|A|^{|\Gamma \cup \Psi_0|} - |A|^{|\Gamma|}\right) \log|\Lambda_n|\,.\end{aligned}$$

Using Lemma 3.13 and the obvious bound $|\Gamma \cup \Psi_0| \le |\Gamma| + |\Psi_0|$, it follows that, eventually almost surely as $n \to \infty$ for all $\Gamma \not\supseteq \Gamma_0$ with $r(\Gamma) \le R_n$,

$$\mathrm{PIC}_\Gamma(x(\Lambda_n)) > \mathrm{PIC}_{(\Gamma \cap \Gamma_0) \cup \Psi_0}(x(\Lambda_n)) - |A|^{|\Gamma|} \left(|A|^{|\Psi_0|} - 1\right) \log|\Lambda_n|\,.$$

Here, by Lemma 3.15,

$$\begin{aligned}&\mathrm{PIC}_{(\Gamma \cap \Gamma_0) \cup \Psi_0}(x(\Lambda_n)) \\ &\qquad > -\log \mathrm{MPL}_{(\Gamma \cap \Gamma_0) \cup \Psi_0}(x(\Lambda_n)) > -\log \mathrm{MPL}_{\Gamma_0}(x(\Lambda_n)) + c\left|\bar{\Lambda}_n\right|,\end{aligned}$$

eventually almost surely as $n \to \infty$ for all Γ as above. This completes the proof, since the conditions $r(\Gamma) \le R_n$ and $|\Lambda_n| \,/\left|\bar{\Lambda}_n\right| \to 1$ imply $|A|^{|\Gamma|} \log|\Lambda_n| = o\left(\left|\bar{\Lambda}_n\right|\right)$. □

3.6 Discussion

A modification of the Bayesian Information Criterion (BIC) called PIC has been introduced for estimating the basic neighborhood of a Markov random field on $\mathbb{Z}^d$, with finite alphabet A. In this criterion, the maximum pseudo-likelihood is used instead of the maximum likelihood, with penalty term $|A|^{|\Gamma|} \log |\Lambda_n|$ for a candidate neighborhood Γ, where Λ_n is the sample region. The minimizer of PIC over candidate neighborhoods, with radius allowed to grow as $o(\log^{\frac{1}{2d}} |\Lambda_n|)$, has been proved to equal the basic neighborhood eventually almost surely, not requiring any prior bound on the size of the latter. This result is unaffected by phase transition and even by non-stationarity of the joint distribution. The same result holds if the penalty term is multiplied by any $c > 0$; the no underestimation part (Proposition 3.11) holds also if $\log |\Lambda_n|$ in the penalty term is replaced by any function of the sample size $|\Lambda_n|$ that goes to infinity as $o(|\Lambda_n|)$.

PIC estimation of the basic neighborhood of a Markov random field is to a certain extent similar to BIC estimation of the order of a Markov chain, and of the context tree of a tree source, also called variable length Markov chain. For context tree estimation via another method see Weinberger, Rissanen and Feder (1995), Bühlmann and Wyner (1999), and via BIC, see Csiszár and Talata (2004). There are, however, also substantial differences. The martingale techniques in Csiszár and Shields (2000) and Csiszár (2002) do not appear to carry over to Markov random fields, and the lack of an analogue of the Krichevsky – Trofimov distribution used in these references is another obstacle. We also note that the "large" boundaries of multidimensional sample regions cause side effects not present in the one dimensional case; to overcome those, we have defined the pseudo-likelihood function based on a window $\bar{\Lambda}_n$ slightly smaller than the whole sample region Λ_n.

For Markov order and context tree estimation via BIC, consistency has been proved by Csiszár and Shields (2000) admitting, for sample size n, all $k \leq n$ as candidate orders, see also Csiszár (2002), respectively by Csiszár and Talata (2004) admitting trees of depth $o(\log n)$ as candidate context trees. In our main result Theorem 3.4, the PIC estimator of the basic neighborhood is defined admitting candidate neighborhoods of radius $o(\log^{\frac{1}{2d}} |\Lambda_n|)$ thus of size $o(\log^{1/2} |\Lambda_n|)$. The mentioned one-dimensional results suggest that this bound on the radius might be relaxed to $o(\log^{1/d} |\Lambda_n|)$, or perhaps dropped completely. This question remains open, even for the case $d = 1$. A positive answer apparently depends on the possibility of strengthening our typicality result Proposition 3.7 to similar strength as the conditional typicality results for Markov chains in Csiszár (2002).

More important than a possible mathematical sharpening of Theorem 3.4, as above, would be to find an algorithm to determine the PIC estimator without actually computing and comparing the PIC values of all candidate neighborhoods. The analogous problem for BIC context tree estimation has been solved: Csiszár and Talata (2004) showed that this BIC estimator can be computed in linear time via an analogue of the "context tree maximizing algorithm" of Willems, Shtarkov, and Tjalkens (1993, 2000). Unfortunately, a similar algorithm for the present problem appears elusive, and it remains open whether our estimator can be computed in a "clever" way.

Finally, we emphasize that the goal of this work was to provide a consistent estimator of the basic neighborhood of a Markov random field. Of course, consistency is only one of the desirable properties of an estimator. To assess the practical performance of this estimator requires further research, such as studying finite sample size properties, robustness against noisy observations, and computability with acceptable complexity.

3.A Appendix

First we indicate how the well-known facts stated in Lemma 3.1 can be formally derived from results in Georgii (1988), using the concepts defined there.

Proof of Lemma 3.1. By Theorem 1.33, the positive one-point specification uniquely determines the specification, which is positive and local on account of the locality of the one-point specification. By Theorem 2.30, this positive local specification determines a unique "gas" potential (if an element of A is distinguished as the zero element). Due to Corollary 2.32, this is a nearest-neighbor potential for a graph with vertex set $\mathbb{Z}^d$ defined there, and Γ_0^i is the same as $B(i)\backslash\{i\}$ in that Corollary. □

The following lemma is a consequence of the global Markov property.

Lemma 3.16. *Let $\Delta \subset \mathbb{Z}^d$ be a finite region with $0 \in \Delta$, and $\Psi = (\cup_{j\in\Delta}\Gamma_0^j) \setminus \Delta$. Then for any neighborhood Γ, the conditional probabilities $Q(a(i)|\, a(\Gamma^i \cup \Psi^i))$ and $Q(a(i)|\, a((\Gamma^i \cap \Delta^i) \cup \Psi^i))$ are equal and translation invariant.*

Proof. Since Δ and Ψ are disjoint, we have

$$\begin{aligned} Q(a(i)|\, a(\Gamma^i \cup \Psi^i)) &= Q\left(a(i)\middle|\, a((\Gamma \cap \Delta)^i \cup (\Psi \cup (\Gamma\backslash\Delta))^i\right) \\ &= \frac{Q\left(a(\{i\} \cup (\Gamma \cap \Delta)^i)|\, a((\Psi \cup (\Gamma\backslash\Delta))^i\right)}{Q\left(a((\Gamma \cap \Delta)^i)|\, a((\Psi \cup (\Gamma\backslash\Delta))^i\right)}\,, \end{aligned}$$

and similarly

$$Q\left(a(i)|\, a((\Gamma^i \cap \Delta^i) \cup \Psi^i)\right) = \frac{Q\left(a(\{i\} \cup (\Gamma \cap \Delta)^i)|\, a(\Psi^i)\right)}{Q\left(a((\Gamma \cap \Delta)^i)|\, a(\Psi^i)\right)}.$$

By the global Markov property, see Lemma 3.1, both the numerators and denominators of these two quotients are equal, and translation invariant. □

The lemma below follows from the definition of Markov neighborhood.

Lemma 3.17. *For a Markov random field with basic neighborhood Γ_0, if a neighborhood Γ satisfies*

$$Q(a(i)|\, a(\Gamma^i)) = Q_{\Gamma_0}(a(i)|\, a(\Gamma_0^i))$$

for all $i \in \mathbb{Z}^d$, then Γ is a Markov neighborhood.

Proof. We have to show that for any $\Delta \supset \Gamma$

$$Q(a(i)|\, a(\Delta^i)) = Q(a(i)|\, a(\Gamma^i)). \tag{3.12}$$

Since Γ_0 is a Markov neighborhood, the condition of the Lemma implies

$$Q(a(i)|\, a(\Gamma^i)) = Q(a(i)|\, a(\Gamma_0^i)) = Q(a(i)|\, a((\Gamma_0 \cup \Delta)^i)).$$

Hence (3.12) follows, because $\Gamma \subseteq \Delta \subseteq \Gamma_0 \cup \Delta$. □

Next, we state two simple probability bounds.

Lemma 3.18. *Let Z_1, Z_2, ... be $\{0,1\}$-valued random variables such that*

$$\text{Prob}\left\{\, Z_j = 1 \mid Z_1, \dots Z_{j-1} \,\right\} \geq p_* > 0, \quad j \geq 1,$$

with probability 1. *Then for any $0 < \nu < 1$*

$$\text{Prob}\left\{ \frac{1}{m} \sum_{j=1}^{m} Z_j < \nu p_* \right\} \leq e^{-m\frac{p_*}{4}(1-\nu)^2}.$$

Proof. This is a direct consequence of Lemmas 2 and 3 in the Appendix of Csiszár (2002). □

Lemma 3.19. *Let Z_1, Z_2, ..., Z_n be i.i.d. random variables with expectation* 0 *and variance D^2. Then the partial sums*

$$S_k = Z_1 + Z_2 + \cdots + Z_k$$

satisfy

$$\text{Prob}\left\{ \max_{1 \leq k \leq n} S_k \geq D\sqrt{n}\,(\mu + 2) \right\} \leq \frac{4}{3}\,\text{Prob}\left\{\, S_n \geq D\sqrt{n}\,\mu \,\right\},$$

moreover if the random variables are bounded, $|Z_i| \le K$*, then*

$$\text{Prob}\left\{ S_n \ge D\sqrt{n}\,\mu \right\} \le 2\exp\left[-\frac{\mu^2}{2\left(1+\frac{\mu K}{2D\sqrt{n}}\right)^2} \right],$$

where $\mu < D\sqrt{n}/K$*.*

Proof. See, for example, in Rényi (1970) Lemma VI.9.1 and Theorem VI.4.1. □

The following three lemmas are of technical nature.

Lemma 3.20. *For disjoint finite regions* $\Phi \subset \mathbb{Z}^d$ *and* $\Delta \subset \mathbb{Z}^d$*, we have*

$$Q(a(\Delta)|\, a(\Phi)) \ge q_{min}^{|\Delta|}.$$

Proof. By induction on $|\Delta|$.

For $\Delta = \{i\}$, $\Xi = \Gamma_0^i \setminus \Phi$, we have

$$\begin{aligned} Q(a(i)|\, a(\Phi)) &= \sum_{a(\Xi)\in A^\Xi} Q(a(i)|\, a(\Phi\cup\Xi))\, Q(a(\Xi)|\, a(\Phi)) \\ &= \sum_{a(\Xi)\in A^\Xi} Q(a(i)|\, a(\Gamma_0^i))\, Q(a(\Xi)|\, a(\Phi)) \ge q_{\min}. \end{aligned}$$

Supposing $Q(a(\Delta)|\, a(\Phi)) \ge q_{\min}^{|\Delta|}$ holds for some Δ, for $\{i\}\cup\Delta$, with $\Xi = \Gamma_0^i \setminus (\Phi\cup\Delta)$, we have

$$\begin{aligned} Q(a(\{i\}\cup\Delta)|\, a(\Phi))) &= \sum_{a(\Xi)\in A^\Xi} Q(a(\{i\}\cup\Delta\cup\Xi)|\, a(\Phi)) \\ &= \sum_{a(\Xi)\in A^\Xi} Q(a(i)|\, a(\Delta\cup\Xi\cup\Phi))\, Q(a(\Delta\cup\Xi)|\, a(\Phi)) \end{aligned}$$

Since $Q(a(i)|\, a(\Delta\cup\Xi\cup\Phi)) = Q(a(i)|\, a(\Gamma_0^i)) \ge q_{\min}$, we can continue as

$$\ge q_{\min}\, Q(a(\Delta)|\, a(\Phi)) \ge q_{\min}^{|\Delta|+1}.$$

□

Lemma 3.21. *The number of all possible blocks appearing in a site and its neighborhood with radius not exceeding* R*, can be upper bounded as follows:*

$$\left|\left\{ a(\Gamma,0) \in A^{\Gamma\cup\{0\}} : r(\Gamma) \le R \right\}\right| \le (|A|^2+1)^{(2R+1)^d/2}.$$

Proof. The number of the neighborhoods with cardinality $m \geq 1$ and radius $r(\Gamma) \leq R$ is

$$\binom{\left((2R+1)^d - 1\right)/2}{m},$$

because the neighborhoods are symmetric. Hence, the number in the proposition is

$$\begin{aligned} |A| + |A| \cdot \sum_{m=1}^{\left((2R+1)^d-1\right)/2} \binom{\left((2R+1)^d - 1\right)/2}{m} |A|^{2m} \\ = |A| \sum_{m=0}^{\left((2R+1)^d-1\right)/2} \binom{\left((2R+1)^d - 1\right)/2}{m} \left(|A|^2\right)^m 1^{\left((2R+1)^d-1\right)/2 - m}. \end{aligned}$$

Now, using the binomial theorem, the assertion follows. □

Lemma 3.22. *Let P and Q be probability distributions on A such that*

$$\max_{a \in A} |P(a) - Q(a)| \leq \frac{\min_{a \in A} Q(a)}{2}.$$

Then

$$\sum_{a \in A} P(a) \log \frac{P(a)}{Q(a)} \leq \frac{1}{\min_{a \in A} Q(a)} \sum_{a \in A} \left(P(a) - Q(a)\right)^2.$$

Proof. This follows from Lemma 4 in the Appendix of Csiszár (2002). □

Bibliography

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statist. Math.* **22** 203–217.

Akaike, H. (1972). Information theory and an extension of the maximum likelihood principle. In *Proceedings of the 2nd International Symposium on Information Theory,* Supplement to Problems of Control and Information Theory (B. N. Petrov and F. Csáki, eds.) 267–281. Akadémia Kiadó, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19** 716–723.

Akaike, H. (1977). On entropy maximization principle. In *Application of Statistics* (P.R. Krishnaiah, ed.) 27–41. North-Holland, Amsterdam.

Anderson, T.W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.* **33** 255–265.

Anderson, T.W. (1963). Determination of the order of dependence in normally distributed time series. In *Time series analysis* (M. Rosenblatt, ed.) 425–446. Wiley, New York.

Azencott, R. (1987). Image analysis and Markov fields. In *Proceedings of the First International Conference on Applied Mathematics, Paris* (J. McKenna and R. Temen, eds.) 53–61. SIAM, Philadelphia.

Baron, D. and Bresler, Y. (2004). An $O(N)$ semipredictive universal encoder via the BWT. *IEEE Trans. Inform. Theory* **50** 928–937.

Barron, A., Rissanen, J. and Yu, B. (1998). The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory* **44** 2743–2760.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *J. Roy. Statist. Soc. Ser. B* **36** 192–236.

Besag, J. (1975). Statistical analysis of non-lattice data. *The Statistician* **24** 179–195.

Bühlmann, P. and Wyner, A.J. (1999). Variable length Markov chains. *Ann. Statist.* **27** 480–513.

Comets, F. (1992). On consistency of a class of estimators for exponential families of Markov random fields on the lattice. *Ann. Statist.* **20** 455–468.

Csiszár, I. (2002). Large-scale typicality of Markov sample paths and consistency of MDL order estimators. *IEEE Trans. Inform. Theory* **48** 1616–1629.

Csiszár, I. and Shields, P.C. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.* **28** 1601–1619.

Csiszár, I. and Talata, Zs. (2004a). Consistent Estimation of the Basic Neighborhood of Markov Random Fields. *Ann. Statist.* Accepted.

Csiszár, I. and Talata, Zs. (2004b). Context Tree Estimation for Not Necessarily Finite Memory Processes, via BIC and MDL. *IEEE Trans. Inform. Theory.* Submitted.

Davisson, L.D. (1965). Prediction error of stationary Gaussian time series of unknown variance. *IEEE Trans. Inform. Theory* **19** 783–795.

Dobrushin, R.L. (1968). The description of a random field by means of conditional probabilities and conditions of its regularity. *Theory Probab. Appl.* **13** 197–224.

Finesso, L. (1992). Estimation of the order of a finite Markov chain. In *Recent Advances in Mathematical Theory of Systems, Control, Networks and Signal Processing, I* (H. Kimura and S. Kodama, eds.) 643–645. Mita Press, Tokyo.

Geman, S. and Graffigne, C. (1987). Markov random fields image models and their applications to computer vision. In *Proceedings of the International Congress Mathematicians* (A. M. Gleason, ed.) **2** 1496–1517. Amer. Math. Soc., Providence, R.I.

Georgii, H.O. (1988). *Gibbs Measures and Phase Transitions.* de Gruyter, Berlin.

GERENCSÉR, L. (1987). Order estimation of stationary Gaussian ARMA processes using Rissanen's complexity. Working paper, Computer and Automation Institute of the Hungarian Academy of Sciences.

GIDAS, B. (1988). Consistency of maximum likelihood and pseudolikelihood estimators for Gibbs distributions. *Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. Math. Appl.* **10** 129–145.

HAMERLY, E.M. and DAVIS, M.H.A. (1989). Strong consistency of the PLS criterion for order determination of autoregressive processes. *Ann. Statist.* **17** 941–946.

HANNAN, E.J. (1980). The estimation of the order of an ARMA process. *Ann. Statist.* **8** 1071–1081.

HANNAN, E.J. and QUINN, B.G. (1979). The determination of the order of an autoregression. *J. Roy. Statist. Soc. Ser. B* **41** 190–195.

HAUGHTON, D. (1988). On the choice of model to fit data from an exponential family. *Ann. Statist.* **16** 342–355.

KRICHEVSKY, R.E. and TROFIMOV, V.K. (1981). The performance of universal encoding. *IEEE Trans. Inform. Theory* **27** 199–207.

MALLOWS, C. (1964). Choosing variables in a linear regression: A graphical aid. Presented at the Central Regional Meeting of the IMS, Manhattan, Kansas.

MALLOWS, C. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MARTÍN, A., SEROUSSI, G. and WEINBERGER, M.J. (2004). Linear time universal coding and time reversal of tree sources via FSM closure. *IEEE Trans. Inform. Theory* **50** 1442–1468.

NEYMAN, J. and PEARSON, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika* **20A** 263–294.

PICKARD, D.K. (1987). Inference for discrete Markov field: The simplest nontrivial case. *J. Amer. Statist. Assoc.* **82** 90–96.

RÉNYI, A. (1970). *Probability Theory.* American Elsevier Publishing Co., Inc., New York.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.

Rissanen, J. (1983a). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11** 416–431.

Rissanen, J. (1983b). A universal data compression system. *IEEE Trans. Inform. Theory* **29** 656–664.

Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific, Singapore.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory* **42** 40–47.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

Shibata, R. (1976). Selection of the order of an autoregressive model by Akaike's information criterion. *Biometrika* **63** 117–126.

Shtarkov, J. (1977). Coding of discrete sources with unknown statistics. In *Topics in Information Theory* (I. Csiszár and P. Elias, eds.) 559–574. North-Holland, Amsterdam.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. Roy. Statist. Soc. Ser. B* **39** 44–47.

Talata, Zs. (2004). Model Selection via Information Criteria. *Period. Math. Hungar.* Invited paper.

Weinberger, M.J., Lempel, A. and Ziv, J. (1992). A sequentional algorithm for the universal coding of finite memory sources. *IEEE Trans. Inform. Theory* **38** 1002–1014.

Weinberger, M.J., Rissanen, J. and Feder, M. (1995). A universal finite memory source. *IEEE Trans. Inform. Theory* **41** 643–652.

Willems, F.M.J. (1998). The context-tree weighting method: Extensions. *IEEE Trans. Inform. Theory* **44** 792–798.

Willems, F.M.J., Shtarkov, Y.M. and Tjalkens, T.J. (1993). The context-tree weighting method: Basic properties. *Tech. Rep., EE Dept., Eindhoven University.* An earlier unabridged version of (Willems, Shtarkov and Tjalkens, 1995).

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

Willems, F.M.J., Shtarkov, Y.M. and Tjalkens, T.J. (2000). Context-tree maximizing. In *Proc. 2000 Conf. Information Sciences and Systems* TP6-7–TP6-12. Princeton, NJ.