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

By

Zsolt Talata

Outline of the 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

1 Introduction

This work is an outline of the Ph.D. dissertation *Model Selection via Information Criteria for Tree Models and Markov Random Fields*. The dissertation consists of three chapters. The first one is a survey of statistical model selection problems. The two other chapters contain the new results of the dissertation: the second chapter is concerned with the estimation of context trees, and the third chapter addresses the model selection problem for Markov random fields.

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.

This outline introduces the model selection problem in Section 2, and summarizes the results in the literature which motivated the new results of the dissertation in Section 3. The two new results are formulated in Sections 4 and 5, indicating the methods of the proofs. The bibliography contains all references of the dissertation.

2 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 $\{ x_i,\, i = 1, \dots, n \}$.

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 $\{ x_t,\, t \in \{t_1, \dots, t_n\} \subset T \}$.

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

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

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:

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

3 Previous results

For various processes it has been proved that the *Bayesian Information Criterion (BIC)* (Schwarz, 1978) and *Minimum Description Length (MDL)* criterion (Rissanen, 1978, 1983a, 1989) lead to strongly consistent estimators of the structure parameter. This means that the minimizer of BIC or MDL over the candidate structure parameters is equal to the true structure parameter, eventually almost surely as the sample size n tends to infinity. 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 $n \geq n_0$.

Consistency proofs are available for, e.g., i.i.d. processes with distributions from exponential families (Haughton, 1988), AR processes (Hannan and Quinn, 1979), ARMA processes (Hannan, 1980), Markov chains (Finesso, 1992) and tree models (Willems, Shtarkov and Tjalkens, 1993, 2000).

All these proofs 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.

For the Markov chain of order k, the Bayesian Information Criterion has the following form:

$$\mathrm{BIC}_k(x_1^n) = -\log \mathrm{ML}_k(x_1^n) + \frac{(|A|-1)|A|^k}{2} \log n,$$

where ML_k denotes the maximum likelihood. Here $(|A|-1)|A|^k$ equals the number of parameters. For the maximum likelihood we have

$$\mathrm{ML}_k(x_1^n) = \prod_{a_1^{k+1} :\, N_n(a_1^{k+1}) \geq 1} \left[\frac{N_n(a_1^{k+1})}{N_n(a_1^k)} \right]^{N_n(a_1^{k+1})},$$

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 MDL criterion can be written as

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

where $P_k^{(n)}$ denotes a coding distribution tailored to the class of Markov chains of order k, and $L(k)$ denotes the codelength of the order k. The usual coding

distributions are the *Krichevsky – Trofimov (KT)* (Krichevsky and Trofimov, 1981) and *Normalized Maximum Likelihood (NML)* distributions.

The explicit form of the KT distribution of order k is

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

The NML distribution is defined as

$$\mathrm{NML}_k^{(n)}(x(n)) = \mathrm{ML}_k(x_1^n) \Big/ \sum_{x_1^n \in A^n} \mathrm{ML}_k(x_1^n) \, .$$

Shtarkov (1977) showed that the sum of maximum likelihoods $\mathrm{ML}_k(x_1^n)$ over all possible samples x_1^n is asymptotically (as n tends to infinity, with k fixed) equal to $(|A| - 1)|A|^k \, 2^{-1} \log n$. Hence, the NML version of the MDL criterion is asymptotically equivalent to the BIC, when the number of candidate orders k is finite. As the following results indicate it, this equivalence does not hold when there is no prior bound on the order k.

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 $0 \leq k < n$ 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, when the coding distribution $P_k^{(n)}$ is either KT_k or $\mathrm{NML}_k^{(n)}$, and the codelength $L(k)$ of the order k satisfies $L(k) = o(k)$, we have

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

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.

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

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:

(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 4.1. *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 4.2. 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 4.1 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 $|\mathcal{T}_0| = 1$.

Example 4.3. *(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. 1. □

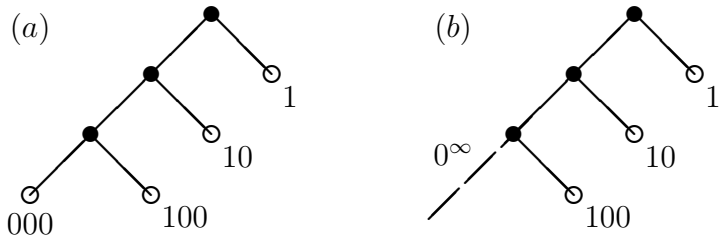

Figure 1: Context tree of a renewal process. (*a*) $k_0 = 3$. (*b*) $k_0 = \infty$.

In the dissertation, 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 (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 $N_n(s)$:

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

Given a sample x_1^n, a *feasible tree* is any tree $\mathcal{T}$ of depth $d(\mathcal{T}) \le \lceil \log n \rceil$ such that $N_n(s) \ge 1$ for all $s \in \mathcal{T}$, and each string s' with $N_n(s') \ge 1$ 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 $N_n(s) \ge r$ 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 4.4. *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 4.5. 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 4.4 involves a slight abuse of terminology. We note that replacing $(|A|-1)/2$ in Definition 4.4 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 4.6. *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 of Theorems 4.6 and 4.9. The proofs of this and the next Theorems are based on the large-scale typicality results on Markov chains (Csiszár, 2000). □

Remark 4.7. Here and in Theorem 4.9 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 code-length (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 4.8. *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 4.9. *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 $n \to \infty$.

Remark 4.10. 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. □

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 dissertation 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 4.11. *The number of computations needed to determine the BIC estimator and the KT estimator in Theorems 4.6 and 4.9 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.*

Theorem 4.12. *Given a sample x_1^n, the number of computations needed to determine the KT estimator in Theorem 4.9 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 4.6 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 4.4 by the smallest member of the sequence $\{k_m\}$ larger than n.

Proof of Theorems 4.11 and 4.12. These Theorems are proved using an extension of the Context Tree Maximizing (CTM) method of Willems, Shtarkov and Tjalkens (1993, 2000). □

5 Consistent Estimation of the Basic Neighborhood of Markov Random Fields

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,

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

The last standard assumption admits unambiguous definition of the conditional probabilities

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

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.2.

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

$$Q(a(i)|\, a(\Delta^i)) = Q(a(i)|\, a(\Gamma^i)) \quad \text{if } \Delta \supset \Gamma,\ 0 \notin \Delta, \tag{2}$$

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 (2) will be called *one-point specification*. All distributions on $A^{\mathbb{Z}^d}$ that satisfy (2) 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.

Lemma 5.1. *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 5.1 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 the dissertation, 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.2. Logarithms are to the base e. Our only assumptions about the sample regions Λ_n will be that

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

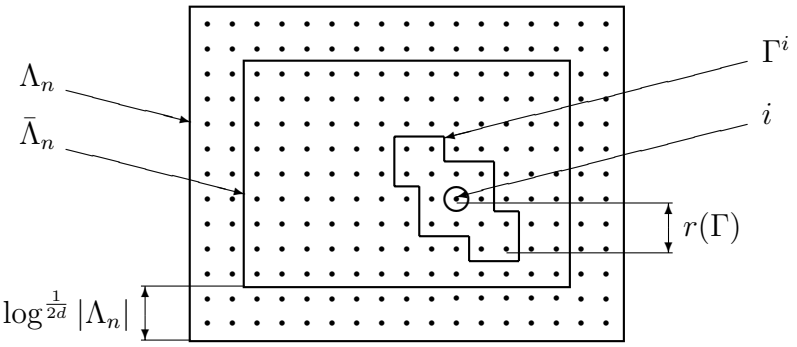

Figure 2: The Γ-neighborhood of the site i, and the sample region Λ_n.

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:

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

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

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

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

Definition 5.2. *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 5.3. *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 $r(\Gamma) \leq r_n$, 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 5.4. *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. The no overestimation part of this Theorem is proved using Besag's "coding technique" (Besag, 1974) and large deviations arguments. The no underestimation part is proved via an entropy argument. □

Remark 5.5. *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. □

6 Bibliography

Publications that contain the parts of the dissertation

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.

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

References of the dissertation

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.

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 non-trivial 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.

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.