

Model Inference and Pattern Discovery by Minimal Representation Method

Jakub Segen and Arthur C. Sanderson

Department of Electrical Engineering
and
The Robotics Institute
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

July 1981

Abstract:

Inference of statistical models and discovery of patterns in random data sets are problems common to many fields of investigation. In particular, in the observation and control of processes where the physical mechanisms are too complex or not well understood to provide a model structure *a priori*, the choice of model structure and model size becomes a key element in the analysis. This paper describes an unsupervised technique for the ranking and model structures and choice of model size based on the expression $[-\log \text{likelihood} + \text{model size (in bits)}]$. This criterion is shown to be equivalent to seeking a parsimonious representation for data, and its derivation is motivated through a Bayesian argument. Limiting properties of the criterion and applications to number of clusters, dimension of a linear predictor, degree of polynomial approximation, or order of a Markov chain are discussed.

This paper was supported in part by the National Science Foundation under grant # ECS-7923893.

SEP 3 1982

629.897

C 23a
82-2
Copy 2

Table of Contents

1 Introduction	1
2 Minimal Representation Criterion	3
3 Properties of $r(q^*, x^n)$	8
4 Practical Implementation	13
5 Applications	14

List of Figures

Figure 1: Universal Turing machine	3
Figure 2: Turing machine as a decoder	7
Figure 3: Identification of an AR model by the minimal representation method and the Akaike criterion. The model generating the data has order 3.	16
Figure 4: Histogram of observations for Example 1.	19
Figure 5: Representation size function vs. the number of clusters in Example 1.	20
Figure 6: Data for clustering in Example 2	21
Figure 7: Representation size evaluation of clustering, Example 2	22

1 Introduction

Although statistical models are now widely used in almost every field of science and engineering the methods of constructing such models from the available data, i.e., the methods of model identification or model inference are still far from being perfected. This paper presents a new approach to model inference which has important advantages over the conventional methods, mainly by being more general and allowing one to construct reliable models of a form that would not be feasible with the classical techniques.

The model inference problem can usually be formulated in the following way. There is a set of models Q and an order (reflexive, transitive and antisymmetric binary relation) $ORD(x^n)$ defined on Q . The order $ORD(x^n)$ is a function of the observations $x^n \triangleq x_1, x_2, \dots, x_n$, such that for any x^n the set Q is well ordered, i.e., Q has a minimal or first element. The model inference procedure finds the minimal element in Q with respect to $ORD(x^n)$, for a given set of observations x^n . The order on the set of models is usually given by an increasing (decreasing) ordering on a set of values of some real function $r(q, x^n)$ of a model q and the observations. This function may be referred to as a measure of goodness of a model with respect to the observations or a measure of fit of a model to the observations. The inference procedure in this case selects a model in Q using a criterion of minimum (maximum) of $r(q, x^n)$.

The most commonly used conventional inference methods are based on either a likelihood or an average quadratic error function. These methods are attractive because they require no knowledge or assumptions about the prior distributions, the resulting estimators have good asymptotic properties and the estimation procedure is usually quite simple. However, those inference methods cannot be meaningfully applied to many interesting and useful classes of models. The problem is usually caused by the fact that certain sets of models are not well ordered with respect to either the likelihood or the quadratic error function. This problem frequently arises in connection with the choice of the dimensionality of a model. Typical examples include the choice of the degree for a polynomial regression, the choice of the order for an autoregressive model or the choice of the number of components for a probability mixture, where the maximum likelihood or the minimum quadratic error criteria always lead to choosing the highest possible dimension.

The conventional ways of treating such problems include interactive methods based on subjective judgment, sometimes supported by an indicator such as an error step size, and analytical techniques. The conventional analytical techniques usually require either a prohibitive amount of computation (e.g. leave-one-out or cross validation method) or an excessive number of data (e.g. holdout method). A survey of these techniques and their various modifications (such as leave-two-out) was provided by Toissant [TOI] with regard to estimation of misclassification probability.

A more recent approach to model identification is represented by the work of Akaike [Akai] and Schwarz [Schwarz]. They incorporated the number of parameters (dimensionality) into the model selection criterion which increasingly penalizes higher dimensional models. The Akaike criterion maximizes the function

$$\log L_j(x^n) - k_j$$

where $L_j(x^n)$ is the likelihood and k_j is the dimension for the model j . The penalty term k_j arose somewhat arbitrarily as an approximate bias of the Kullback information function. The criterion of Schwarz maximizes

$$\log L_j(x^n) - (1/2)k_j \log n$$

and the resulting estimates were shown to be asymptotically optimal under a 0-1 loss function, for i.i.d. observations and for a specific class of linear models. Neither of these methods, however, provides a basis for discriminating among models having parameters of different type, range or different effect on the likelihood function since only the total number of the parameters appears in the criterion.

These deficiencies in the above methods stimulated the search for a more general inference technique. The result of this investigation, a criterion selecting a model that leads to the most compact representation of the observations, was initially introduced in [Minimum]. This criterion is motivated by a Bayesian argument with a 0-1 loss function and a prior distribution assigning lower probability values to more complex models. Such choice of the prior distribution was stimulated by the ideas contained in the work of Solomonoff [Solomonoff1] on inductive inference. Solomonoff has also considered the model inference problem [Solomonoff1], [Solomonoff2] in a form of a general probability estimation. The approach that he proposes, however, requires two impossible things: To find the shortest program for a Turing machine for generating a given sequence, and to calculate a sum of an infinite series without knowing the analytical expression for the elements. The first problem is in general unsolvable [Chaitin1] and the solution to the second task can only be approximated. Although Solomonoff's approach to the probability estimation problem is practically and theoretically unfeasible in its direct form, his ideas initiated the development of algorithmic information theory [Kolmogorov] [Chaitin], [Willis] to which our method is to some extent related.

An approach resulting in conclusions that are equivalent to ours but based on a different motivation has been proposed earlier by Rissanen [R11]. His method uses Gibbs' theorem as a basis for model identification, although he also noticed the possibility of a Bayesian interpretation. Rissanen has also shown that this method applied to AR and ARMA models with an unspecified order results in consistent order estimates [R11], [R12]. This makes it better than the Akaike method which cannot estimate consistently the order of AR model as was shown by Shibata [Shibata].

The methods proposed in [Minimum], and in [R11] are also related to the work of Wallace and Boulton [Boulton]. They explored an idea of using the minimum of a required storage size as a criterion in the context of classification.

This paper expands the approach presented in [Minimum] and introduces some new results on asymptotic properties of the method. Section 2 provides the motivation for the proposed inference procedure. Section 3 presents some of its properties and Section 4 illustrates the procedure with several applications.

2 Minimal Representation Criterion

The problem of evaluating and comparing statistical models will be treated as a special case of a more general problem described below.

Assume that we have a universal Turing machine T with three tapes: unidirectional read-only input tape, bidirectional read-write work tape and unidirectional write-only output tape, represented schematically in Fig. TUR1. The machine T reads a binary program p from the input tape and writes the output string $T(p)$ on the output tape. All finite binary strings are assumed to be legal as programs for T . A given program p may produce either a finite or an infinite output string $T(p)$, or no output at all (then $T(p)$ is a null string). After producing a finite output the machine may halt or it may proceed forever.

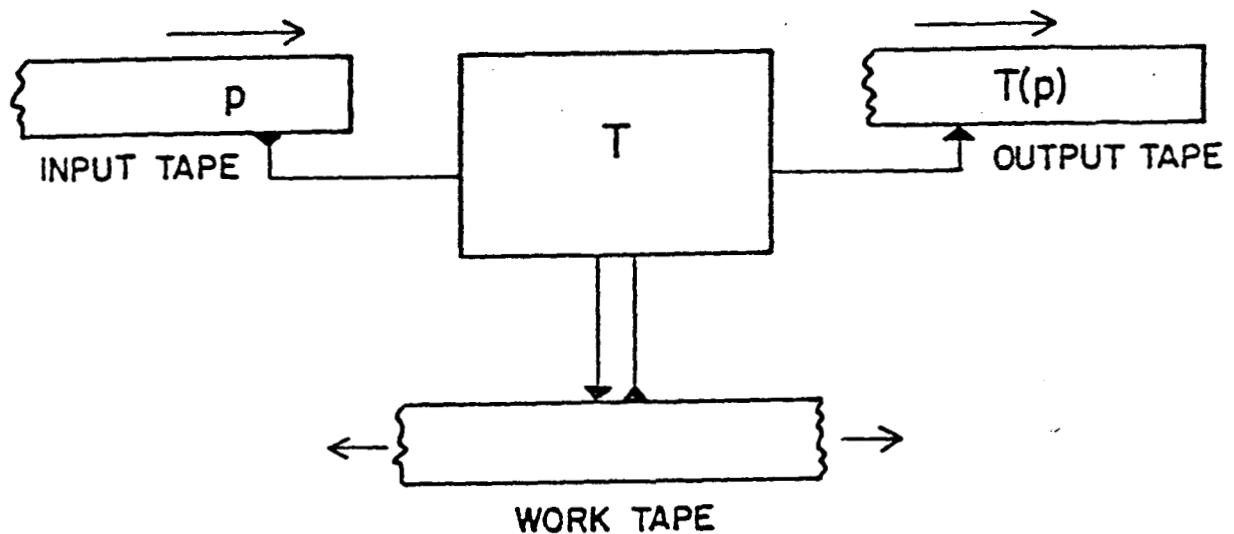


Figure 1: Universal Turing machine

If there are no specific assumptions about the programs then it is reasonable to choose as the a priori probability for a given program p the value

$$P(p) = 2^{-s(p)}$$

where $s(p)$ is the length of the program p in bits.

This choice is equivalent to an assumption that the program p is generated by $s(p)$ random tosses of a fair coin, and it may be considered to be an extension of the principle of insufficient reason [Solomonoff1].

For a given sequence of integer observations

$$x^n \triangleq x_1, x_2, \dots, x_n$$

we will call a program p a *representation* of x^n if the sequence x^n is a prefix of the output string produced by p . Define a *class of representations* of x^n as

$$R(x^n) = \{p \mid x^n = \text{PREFIX}[T(p)]\}$$

Similarly, for a given set H of programs, a *class of representations of x^n in H* can be defined as

$$R(x^n|H) \triangleq \{p \mid x^n = \text{PREFIX}[T(p)], p \in H\} = R(x^n) \cap H$$

Given a program p , the conditional probability that x^n is a prefix of $T(p)$ is

$$P(x^n|p) = \begin{cases} 1 & \text{if } p \in R(x^n) \\ 0 & \text{otherwise} \end{cases}$$

Hence, the posterior probability of p , given that x^n is a prefix of $T(p)$ is

$$P(p|x^n) \propto \begin{cases} 2^{-s(p)} & \text{if } p \in R(x^n) \\ 0 & \text{otherwise} \end{cases}$$

where the symbol \propto means "is proportional to".

Similarly, for programs from a given set H

$$P(p|p \in H) = \frac{P(p \cap H)}{P(H)} = \begin{cases} 2^{-s(p)}/P(H) & \text{if } p \in H \\ 0 & \text{otherwise} \end{cases}$$

and

$$P(p|x^n, p \in H) \propto \begin{cases} 2^{-s(p)} & \text{if } p \in R(x^n|H) \\ 0 & \text{otherwise} \end{cases}$$

For a given sequence x^n and a set of programs H one may attempt to find a program p^* in H which maximizes the a posteriori probability $P(p|x^n, p \in H)$. It follows from the above formula that this program is the minimal length representation of x^n in H , i.e.

$$s(p^*) = \min \{s(p) \mid p \in R(x^n|H)\}$$

The *minimal representation criterion* is a rule selecting the minimal length representation of a sequence x^n in a given set of programs as the most likely program producing x^n .

The minimal representation criterion can be used to discriminate among competing statistical models as will be shown below.

Definition

A function $f: I^* \rightarrow [0,1]$ is *e-computable* if there exists a computable function $f_c(x^n, k)$ which for any $x^n \in I^*$ and for any $k > 0$ evaluates the first k digits of a binary expansion of $f(x^n)$ if $f(x^n) > 0$, or it returns some predefined code if $f(x^n) = 0$.

Lemma 1

Let C be a class of discrete probability distributions for x^n such that each distribution $P_q(\cdot)$ can be uniquely determined by a discrete parameter set q , $q \in Q$, and for any x^n there is a q such that $P_q(x^n) > 0$, and $P_q(x^n)$ is e-computable as a function of q and x^n .

There exists a nonempty set of programs $H(C)$ and a computable isomorphism

$$v: \{(n, q, x^n) \mid P_q(x^n) > 0, q \in Q\} \rightarrow H(C)$$

such that if $v(n, q, x^n) = p$ then

$$s(p) = D + s(n) + s(q) + \lceil -\log P_q(x^n) \rceil,$$

where D is a constant independent of x^n and the notation $\lceil y \rceil$ means the smallest integer greater than or equal to y . The inverse of v is also computable.

Proof

First, notice that if $P_q(x^n)$ is e-computable then the function

$$h(q, x^n) = \begin{cases} \lceil -\log P_q(x^n) \rceil, & \text{if } P_q(x^n) > 0 \\ -1, & \text{if } P_q(x^n) = 0 \end{cases}$$

is computable, since to find $\lceil -\log y \rceil$ for $y > 0$ it is sufficient to evaluate y up to the first nonzero digit. For any q and any n the sequences x^n having nonzero probability $P_q(x^n)$ can be encoded by binary strings forming a prefix set, in such a way that if $c_q(x^n)$ is a code for x^n then

$$s[c_q(x^n)] = \lceil -\log P_q(x^n) \rceil$$

and there are effective encoding and decoding procedures. To prove the above statement we present the procedures for encoding and decoding.

Let $y^n(1), y^n(2), y^n(3) \dots$ be the natural enumeration of the sequences of length n .

Encoding procedure E : To find a code for a sequence x^n , given q , compute $h(q, y^n(i))$ for all the sequences $y^n(i)$ which precede x^n in the natural enumeration. Then, to each of the sequences $y^n(i)$ having a nonzero probability assign as a code $c(i)$ the smallest binary number of length $h(q, y^n(i))$ which is not a prefix or an extension of the codes $c(1), c(2), \dots, c(i-1)$. The correctness of this procedure is implied by Theorem 3.2 of Chaitin [Chaitin].

Decoding procedure E^{-1} : To decode a code c , given n and q , compute for $i = 1, 2, \dots$ the codes $c(y^n(i))$ and the values

$$z(i) = \lceil -\log [1 - P_q(y^n(1)) - P_q(y^n(2)) - \dots - P_q(y^n(i))] \rceil$$

until either of the following two conditions is satisfied:

1. $c = c(y^n(i))$; then return $y^n(i)$
2. $z(i) > s(c)$; then halt (error)

The second condition insures that the procedure will terminate even if c is not a code of any sequence.

Now, the program $p = v(n, q, x^n)$ will contain the number n , the parameter set q , the code $c_q(x^n)$, the decoding program E^{-1} , and an executive routine (EXEC) placed in first position on the input tape:

$$p = \{\text{EXEC}, E^{-1}, n, q, c_q(x^n)\}.$$

The routine EXEC is a program for T that reads the remaining portion of p from the input tape, executes the procedure E^{-1} , and writes the result on the output tape (see Fig. TUR2). We can see that

$$T[v(n, q, x^n)] = x^n$$

Clearly, for a given class of probability functions C one can construct the mapping v as a procedure producing a program p in the above form from n, q , and x^n . This procedure will use the encoding routine E to compute $c_q(x^n)$.

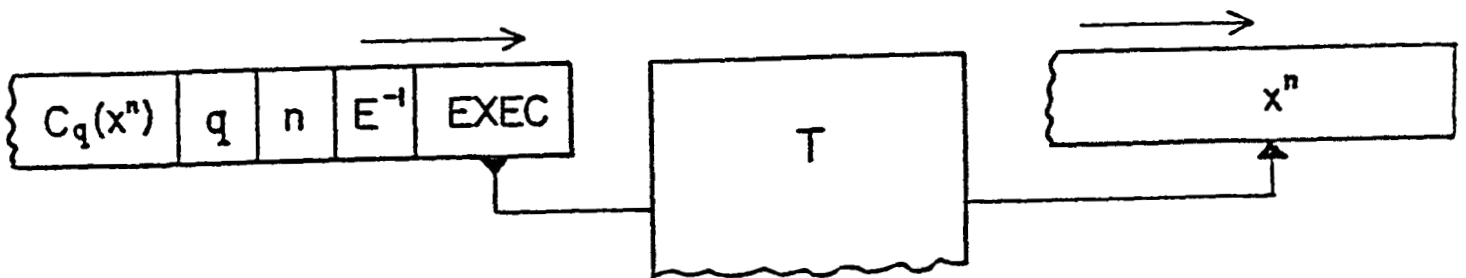


Figure 2: Turing machine as a decoder

The inverse mapping v^{-1} is a procedure which simply executes the program p on the machine T and in addition copies n and q from p , returning (n, q, x^n) .

The set $H(C)$ is given as

$$H(C) = \{ p \mid v(n, q, x^n) = p, P_q(x^n) > 0, q \in Q \}$$

and is clearly nonempty.

The length of the program p corresponding to given n, q , and x^n is

$$s[v(n, q, x^n)] = s(\text{EXEC}) + s(E^{-1}) + s(n) + s(q) + s(c_q(x^n))$$

$$= D + s(n) + s(q) + \lceil -\log P_q(x^n) \rceil$$

and this completes the proof of the lemma.

The above lemma shows a correspondence between a class of probability distributions C and a class of programs $H(C)$. The statistical inference problem can be now approached by applying the minimal representation criterion to find the most likely program within the set $H(C)$ which produces the observed sequence. To use this approach it is not necessary to construct the set $H(C)$. We can see that for a given sequence x^n

$$\min_{p \in H(C)} s(p) = D + s(n) + \min_{q \in Q} \{s(q) + \lceil -\log P_q(x^n) \rceil\}$$

So, the probability distribution corresponding to the minimal length representation of the sequence x^n will be determined by q^* such that

$$s(q^*) + \lceil -\log P_{q^*}(x^n) \rceil = \min_{q \in Q} \{s(q) + \lceil -\log P_q(x^n) \rceil\}$$

To discriminate among probability distributions we will minimize the function:

$$r(q, x^n) = s(q) - \log P_q(x^n)$$

since if

$$r(q^*, x^n) = \min_{q \in Q} r(q, x^n)$$

then

$$s(q^*) + \lceil -\log P_{q^*}(x^n) \rceil = \min_{q \in Q} \{s(q) + \lceil -\log P_q(x^n) \rceil\}$$

It will be shown in the next section that such q^* always exists.

One may notice that if the a priori probability of q is taken as $P(q) = 2^{-s(q)}$ then $r(q, x^n) \propto P(q|x^n)$, so minimizing $r(q, x^n)$ is equivalent to seeking maximum of the a posterior probability of q , given observation sequence x^n . However, by the former derivation, the principle of minimum of $r(q, x^n)$ is shown to be a special case of a more general principle- the minimal representation criterion.

3 Properties of $r(q^*, x^n)$

As mentioned in the introduction, the likelihood function does not always have a maximum in a given set of probability measures. In contrast, the function $r(q, x^n)$ not only reaches the minimum, but the minimum of $r(q, x^n)$ can always be found in a finite time.

Lemma 2

There is an effective procedure for finding q^* such that

$$r(q^*, x^n) = \min_{q \in Q} r(q, x^n)$$

Proof

The search procedure can proceed as follows: First find some value q' , such that $P_{q'}(x^n) > 0$. This

can be effectively done by enumeration, since it has been assumed (in the definition of the class C) that such q' always exists for a given x^n .

Clearly, $r(q, x^n) \geq s(q)$, so $s(q) > r(q', x^n)$ implies that $r(q, x^n) > r(q', x^n)$. Therefore, all the values q such that $r(q, x^n) \leq r(q', x^n)$ will be in a finite set

$$Q' = \{q | s(q) \leq r(q', x^n)\}.$$

The function $r(q, x^n)$ has a minimum in Q' which is also a minimum in Q , and it can be found in a finite time, which proves the lemma.

One property of an estimator that statisticians usually attempt to show is consistency, i.e., convergence of the estimated probability distribution to the true distribution when the number of observations tends to infinity. Strong consistency of the estimates of AR and ARMA models obtained through a procedure equivalent to minimizing $r(q, x^n)$ was shown by Rissanen [R12], [R13].

In general, to speak of consistency in the usual sense, i.e., mean square consistency, consistency in probability, or strong consistency, one has to assume that the true probability distribution lies within the closure of a given set of distributions. This assumption, however, seems to be unnecessarily restrictive for at least two reasons:

1. Usually, one does not know the general form of the true distribution for a given set of observations.
2. Even if the general type of the true distribution is known it may often be advantageous to use a different form of the modeling distribution (for example, a distribution which is simpler or easier to identify than the true one, and still adequate with respect to some objective such as data prediction or compression).

Therefore, to characterize the limiting behavior of the estimates an alternative form of consistency will be studied; a convergence of the estimates to a certain best model of the true distribution.

Definition

A probability distribution $P_{q^*}(x^n)$ is a *best model* of $P(x^n)$ if the function

$$l(q) \triangleq \lim_{n \rightarrow \infty} E[\log P(x_n | x^{n-1}) - \log P_q(x_n | x^{n-1})]$$

is defined on some subset Q' of Q , and it has its minimum at q^* , i.e., $\min_{q \in Q'} l(q) = l(q^*)$.

There may be more than one best model within a given class of distributions, so we will talk about a set of best models B^* . The above definition is motivated by the following properties of $l(q)$.

Lemma 3

If the function $l(q)$ is defined, then:

1. $l(q) \geq 0$
2. If for some q there is $m < \infty$ such that $P(x^n) = P_q(x^n)$ a.s., for all $n \geq m$, then $l(q) = 0$.

3. If the process defined by $P_q(x^n)$ is stationary and r -order Markov and $P(x^n)$ defines a stationary process, then $I(q) = 0$ implies that for every $n > r$

$$P(x_n|x^{n-1}) = P_q(x_n|x^{n-1}) \text{ a.s.}$$

Proof

The properties 1 and 2 are implied directly by the properties of the Kullback information [Kull].

Property 3: Let

$$H(x_n|x^{n-1}) \triangleq -E \log P(x_n|x^{n-1}),$$

and

$$H_q \triangleq -E \log P_q(x_n|x^{n-1}) \text{ for } n > r.$$

The value H_q is well defined under the conditions stated in 3. The condition $I(q) = 0$ is equivalent to

$$\lim_{n \rightarrow \infty} [H_q - H(x_n|x^{n-1})] = 0$$

and it implies that $H_q = H(x_n|x^{n-1})$ for $n > r$, since

$H_q - H(x_n|x^{n-1}) \geq 0$, for $n > r$, by Kullback information properties [Kull]

and $H(x_n|x^{n-1})$ is nonincreasing.

Hence, for every $n > r$

$$E \log [P(x_n|x^{n-1}) / P_q(x_n|x^{n-1})] = 0$$

which proves the property 3.

If one cannot assume that the true distribution is approachable by the estimator, it may be interesting whether best models exist and if so, how the estimates behave with respect to the class of the best models B^* . Certain cases when something can be said about the existence of the best models and the behaviour of the minimal representation estimates with respect to B^* are presented by the following three lemmas.

Lemma 4

If the following assumptions are satisfied,

1. The true model is stationary, ergodic and it has finite entropy.
2. The models P_q are stationary and finite order Markovian.
3. For some $M < \infty$ and for all $q \in Q$, $s(q) \leq M$.

then

- (i) The set of best models B^* is not empty
- (ii) $P\{q_n^* \in B^* \text{ for all but finitely many cases, as } n \rightarrow \infty\} = 1$

i.e., the minimal representation estimator will almost surely select a best model every time, after some finite number of steps.

Proof

First, notice that assumption 3 implies that the number of models is finite so their order is bounded by some $k < \infty$.

(i) The initial assumption that for any x^n there is q such that $P_q(x^n) > 0$ implies that there is q' such that $P_{q'}(x_n|x^{n-1}) > 0$ for any $n > k$. This can be shown by contradiction. Namely, if the above statement is not true, then for any $q \in Q$ there is $k+1$ element sequence: a_0, a_1, \dots, a_{k+1} , such that $P_q(a_0|a_1, a_2, \dots, a_{k+1}) = 0$. If one constructs x^n as a concatenation of all such sequences then for any q , $P_q(x^n) = 0$ which contradicts the initial assumption. Hence, for some $q' \in Q$, $P_{q'}(x_n|x^{n-1}) > 0$ for all $n > k$, so $E \log P_{q'}(x_n|x^{n-1})$ is finite, and by assumptions 1 and 2 the limit $I(q')$ is defined and finite. Therefore, the subset Q' is nonempty and finite, since Q is finite, so $I(q)$ has a minimum in Q' and the class of best models B^- is not empty.

(ii) The elements of the infinite sequence $\{q_n^*\}$ belong to a finite set $\{q|s(q) \leq M\}$, so there is a nonempty set $Q'' \subseteq Q$, such that each $q \in Q''$ occurs in $\{q_n^*\}$ infinitely many times, and for some finite m $q_n^* \in Q''$ for all $n \geq m$.

We will prove that $P(Q'' \subseteq B^-) = 1$ by showing that $P(Q'' \not\subseteq B^-) = 0$. If $Q'' \not\subseteq B^-$, then there is $q' \in Q''$ which is not a member of B^- . First, we consider the case when there is a sequence of values, a_0, a_1, \dots, a_k , such that $P(a_0, a_1, \dots, a_k) > 0$ and $P_{q'}(a_0|a_1, a_2, \dots, a_k) = 0$. If this sequence occurs in x^N , then $r(q', x^n) = \infty$ for all $n \geq N$, so $q' \neq q_n^*$ for all $n \geq N$ since for each x^n there is q such that $r(q, x^n) < \infty$. The element q' can occur in $\{q_n^*\}$ infinitely many times only if the sequence a_0, a_1, \dots, a_k never occurs in $\{x_n\}$, but this event has probability 0.

Now, assume that $P(a_0, a_1, \dots, a_k) > 0$ always implies that $P_{q'}(a_0|a_1, a_2, \dots, a_k) > 0$. The value

$$H_{q'} = -E \log P_{q'}(x_n|x^{n-1})$$

is then finite for all $n > k$, and the limit

$$I(q') = H_{q'} - \lim_{n \rightarrow \infty} H(x_n|x^{n-1})$$

exists. If $q' \notin B^-$, then there exists $q'' \in B^-$ such that $I(q'') < I(q')$, so $H_{q''} < H_{q'}$, giving

$$H_{q'} - H_{q''} = E[\log P_{q''}(x_n|x^{n-1}) - \log P_{q'}(x_n|x^{n-1})] = \epsilon > 0$$

for all $n > k$.

Now, let

$$\begin{aligned} R_n(q', q'') &\triangleq (1/n) * [\log P_{q''}(x^n) - \log P_{q'}(x^n)] \\ &= (1/n) * \sum_{i=k+1}^n [\log P_{q''}(x_i|x^{i-1}) - \log P_{q'}(x_i|x^{i-1})] + C/n. \end{aligned}$$

It follows from assumptions 1 and 2 that

$$\lim_{n \rightarrow \infty} R_n(q', q'') = \text{a.s. } H_{q'} - H_{q''} = \epsilon$$

so almost surely there exists N such that $R_n(q', q'') > \epsilon/2$ for all $n > N$, hence

$$\log P_{q''}(x^n) - \log P_{q'}(x^n) > n * \epsilon/2 \quad \text{for all } n > N, \text{ a.s.}$$

So there is $N' < \infty$, such that

$$\log P_{q''}(x^n) - \log P_{q'}(x^n) > s(q'') - S(q') \quad \text{for } n > N', \text{ a.s.}$$

Hence, $r(q', x^n) > r(q'', x^n)$ for all $n > N'$, a.s. which implies that

$$P\{q_n^* = q', \text{ infinitely often (i.o.) as } n \rightarrow \infty\} = 0.$$

Therefore, $Q'' \not\subseteq B^{\sim}$ always implies a zero probability event, so

$$P(Q'' \not\subseteq B^{\sim}) = 0$$

and

$$P(Q'' \subseteq B^{\sim}) = 1$$

which proves (ii).

If the size of the models is unbounded, then the set of best models might be empty. It may happen either if the function $l(q)$ is undefined on the entire set Q , or if the function $l(q)$ is defined on some infinite subset Q' of Q and for every $q \in Q'$ there is $q' \in Q'$ such that $l(q') < l(q)$. We will show that if B^{\sim} is an empty set, then the size of the minimal representation estimates will almost surely tend to infinity.

Lemma 5

If the assumptions 1 and 2 of Lemma 4 are satisfied and B^{\sim} is empty, then

$$P\{s(q_n^*) \rightarrow \infty\} = 1$$

Proof

If $s(q_n^*)$ does not tend to infinity, then $s(q_n^*) \leq M$ infinitely often as $n \rightarrow \infty$, for some finite M . This implies (as in the proof of Lemma 4) that there is q such that $q_n^* = q$ infinitely often as $n \rightarrow \infty$. The set B^{\sim} is empty, so either $l(q)$ is defined and there is q' such that $l(q') < l(q)$, or $l(q)$ is undefined which, under assumptions 1 and 2, can happen only if there is an event a_0, a_1, \dots, a_k , such that $P(a_0, a_1, \dots, a_k) > 0$ and $P_q(a_0 | a_1, a_2, \dots, a_k) = 0$. It has been shown in the proof of Lemma 4, (ii) that in either of these cases

$$P\{q_n^* = q, \text{ i.o. as } n \rightarrow \infty\} = 0.$$

Hence the event $\text{NOT}\{s(q_n^*) \rightarrow \infty\}$ implies a probability zero event, so

$$\begin{aligned} P\{\text{NOT}[s(q_n^*) \rightarrow \infty]\} &= 0, \text{ and} \\ P\{s(q_n^*) \rightarrow \infty\} &= 1. \end{aligned}$$

If we do not know whether the set B^{\sim} is empty or not, but we know that the minimal representation estimator produces a sequence of estimates whose size is bounded, how could this knowledge be utilized? The following Lemma provides an answer.

Lemma 6

If the true model is randomly chosen from a class of stationary and ergodic models, and the models P_q are stationary, ergodic and finite order Markovian, then

$$P\{s(q_n^*) \leq M \text{ implies that } B^{\sim} \text{ is not empty and } q_n^* \in B^{\sim}, \text{ i.o. as } n \rightarrow \infty\} = 1.$$

Proof

Lemma 5 implies that $P\{s(q_n^*) \leq M \text{ and } B^{\sim} \text{ is empty}\} = 0$. Also, from the proof of Lemma 4(ii) it is evident that

$P\{s(q_n^*) \leq m \text{ and } q_n^* \notin B^{\sim}, \text{ i.o. as } n \rightarrow \infty\} = 0.$

By taking a sum of the above events we obtain

$P\{s(q_n^*) \leq M \text{ and } [B^{\sim} \text{ is empty or } q_n^* \in B, \text{ i.o. as } n \rightarrow \infty]\} = 0$

and since the above event is a negation of the event in the thesis, this proves the lemma.

Lemma 6 shows that from knowledge that size of minimum representation estimates is bounded one can almost surely conclude that a best model exists and the estimates converge to the set of best models. This result can be used in practice to evaluate a given class of models used to explain the observed data. If the size of the minimal representation estimates reaches a stable level and then remains constant with the introduction of new observations, then the assumed class of models may be considered to be adequate. Obviously, one cannot guarantee that the model size will remain constant forever, but as long as it does it corroborates the hypothesis of existence of a best model and of convergence of the inference process.

It might be interesting to notice the relation of the above results to the intuitive approach used in scientific inference, which preserves hypotheses leading to stable models, and calls for a search of an alternative hypothesis whenever the current hypothesis yields models whose complexity continues to increase with the introduction of new data [Kuhn].

4 Practical Implementation

The minimal representation criterion can be used with at most countably infinite classes of probability distributions. This is not a major limitation for the parametric distributions, since any reasonably well behaved function of real parameters can be arbitrarily closely approximated by a function of rational parameters as is done in numerical statistics.

To apply the criterion it is necessary to specify a form of representation for the parameters q , to evaluate $s(q)$. While the choice of this representation may to some extent affect the estimation results for a given set of observations, the limiting properties of the minimal representation estimator, shown in the preceding section, do not depend on this choice.

Clearly, any chosen scheme for representing the parameters should allow one to express all the members of the set Q . Furthermore, the parameter representation should be least redundant in the sense that given two representation schemes 1 and 2, such that $s_1(q) < s_2(q)$ for some members of Q , and $S_1(q) = s_2(q)$ for the remaining elements of Q , then the scheme 1 should be preferred over the scheme 2.

The following representation schemes can be used for typical applications:

- If Q is a finite set with m elements, then its members are represented by $\lceil \log m \rceil$ bit numbers. So, $s(q) = \lceil \log m \rceil$ for every q . Natural number n is represented in a ternary code using 2 bits for each ternary digit: (00) = 0, (01) = 1, (10) = 2; the remaining code (11) indicates the end of the number.

$$s(n) = 2 \cdot \lceil \log_3(n+1) \rceil + 2$$

- Integer i is a natural number with a one bit sign, so

$$s(i) = 2 \cdot \lceil \log_3(|i| + 1) \rceil + 3$$

- Rational number F having a finite binary expansion is represented in a normalized floating point form as a pair of integers i and j , where

$$i = \lfloor \log_2(|F| + 1) \rfloor$$

$$F = c \cdot 2^i$$

$$c = j \cdot 2^{-\lfloor \log_2|j| + 1 \rfloor}$$

and $|j|$ is the smallest integer satisfying the above.

$$s(F) = s(i) + s(j)$$

Notice that $s(j) = 2 \cdot \lceil d(F)/\log_2 3 \rceil + 3$ where $d(F)$ is the number of significant binary digits in F or in its characteristic \bar{C} .

The application of the minimal representation criterion to a problem where the probability distribution is a function of integer and real parameters will involve a search over values of the parameters and over the precisions of binary expansions of the real parameters. A side result of this process is that an estimate of each real parameter is given in a precision which is in certain sense optimal.

An exhaustive search can usually be avoided if the set Q of models can be represented as a sum of sets Q_i such that the maximum likelihood or an equivalent estimation procedure can be applied to any Q_i , and within any set Q_i the value $s(q)$ depends only on the precisions of the parameters. If the probability distribution is reasonably well behaved as a function of real parameters, then a near optimal value of $r(q, x^n)$ within each set Q_i can be found by calculating the maximum likelihood estimates with the highest allowable precision and then searching over their possible truncations to minimize $r(q, x^n)$. The global minimum of $r(q, x^n)$ in Q is then found by comparing the estimates obtained from each of the sets Q_i . An alternative procedure, which has been used by Boulton and Wallace [Boulton] and Rissanen [R11], is based on the assumption that the truncation errors are uniformly distributed, and it determines the optimal truncation levels analytically.

5 Applications

Several examples presented below illustrate the applicability of minimal representation approach to inference problems where the maximum likelihood method would fail.

1. Autoregressive model

The autoregressive model of order k for time series has a form:

$$x_n = a_1 x_{n-1} + a_2 x_{n-2} + \dots + a_k x_{n-k} + a_n$$

where a_i , $i = 1, 2, \dots, k$ are constant coefficients, and $\{u_n\}$ is a discrete white noise with variance σ^2 . If

the order k is fixed, then the maximum likelihood method may be used to estimate the coefficients a_i and the variance σ^2 [Box], [And]. However, if the order is unknown and has to be determined, the maximum likelihood method is not sufficient. To simplify the expression for $P(x^n)$ we assume that $x_n = a_n$ for $n = 1, 2, \dots, k$.

For $n > k$

$$P_q(x_n | x^{n-1}) = (2\pi\sigma^2)^{-1/2} \cdot e^{-v^2/2\sigma^2} \cdot D$$

where $v^n = x_n - a_1 x_{n-1} - a_2 x_{n-2} - \dots - a_k x_{n-k}$ and D is a constant normalizing for finite precision of x_n . So

$$-\log P_q(x^n) = n \cdot \log(2\pi\sigma^2)/2 + (\log e/2\sigma^2) \cdot \left[\sum_{i=k+1}^n v_i^2 + \sum_{i=1}^k x_i^2 \right] + n \log D$$

and $S(q) = S(a_1) + S(a_2) + \dots + S(a_k) + S(\sigma) + S(k)$. The inference procedure searches for a set of coefficients and variance that minimize the value of $r(q, x^n) = S(q) - \log P_q(x^n)$. The practical algorithm which has been implemented uses the method described in [Kashyap] to estimate the coefficients and the variance for a given order k , then it iteratively truncates the parameters to minimize $r(q, x^n)$ for that order, and finally it selects the order that gives a global minimum of $r(q, x^n)$.

Example

The AR model identification algorithm based on the minimal representation method has been applied to sequences of 50, 100, 200 and 400 values generated by a 3rd order autoregressive model with coefficients: $a_1 = 0.7$, $a_2 = -0.5$, and $a_3 = 0.5$. For a comparison, a procedure using Akaike criterion was applied to the same data.

The minimal representation procedure chose a 3rd order model in each case (50, 100, 200 and 400 observations). The Akaike method selected the correct order only for 200 and 400 observations, while it selected a 12th order models for 50 and 100 observations. The behavior of both criterion functions vs. the order of an AR model for each of the cases is shown in Figure AR. Each function was subject to a transformation $y = \log(x - \min x + 1)$ to show more clearly the position of its minimum.

2. Polynomial fitting

We are trying to find a polynomial function to represent a relationship between the variables x and y . The measurements of y are taken at predetermined values of the variable x . The assumed model of the relationship has a form:

$$y_n = a_0 + a_1 x_n + a_2 x_n^2 + \dots + a_k x_n^k + u_n$$

where $\{u_n\}$ is a Gaussian white noise with variance σ^2 . For n measurements

$$\log P_q(y^n) = n \log(2\pi\sigma^2)^{-1/2} + (\log e/2\sigma^2) \cdot \sum_{i=1}^n v_i^2 + n \log D$$

where $v_i = y_n - a_0 - a_1 x_n - \dots - a_k x_n^k$, and

$$S(q) = S(a_0) + S(a_1) + \dots + S(a_k) + S(\sigma^2) + S(k)$$

The algorithm for finding near-minimum of $r(q, x^n)$ in this case can be constructed similarly to the previous one, by using the conventional least squares method for estimating the coefficient at a given

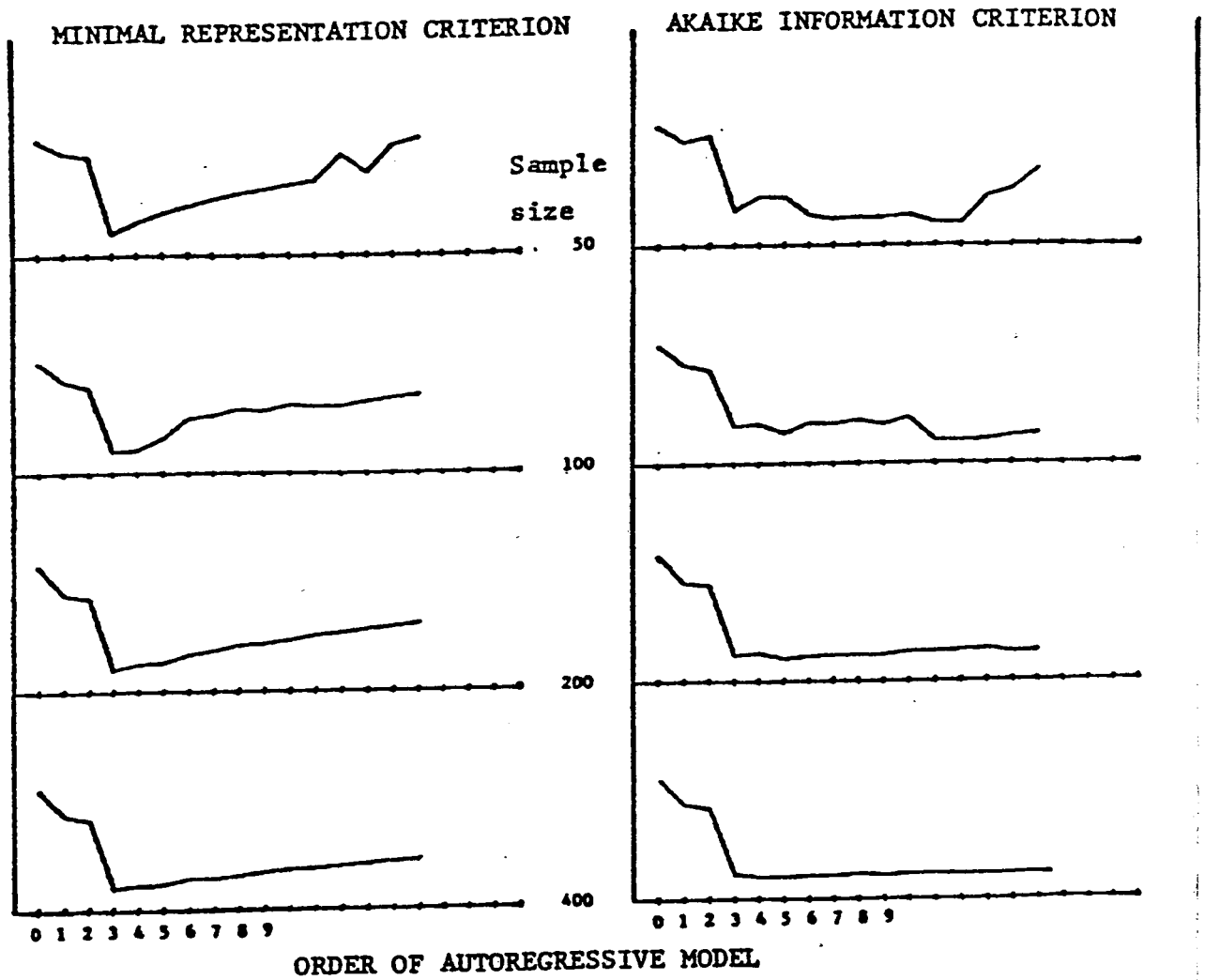


Figure 3: Identification of an AR model by the minimal representation method and the Akaike criterion. The model generating the data has order 3.

polynomial degree k , iterating over the parameter precision and selecting the polynomial degree that gives the global minimum of $r(q, x^n)$. This method can possibly be extended to a procedure for fitting spline functions.

3. Clustering

The term clustering generally refers to the grouping of a given set of objects into subsets containing objects with similar properties. If the probability distributions within clusters and the distribution of the clusters are specified in their general form, then the clustering may be considered and used as an identification procedure for a statistical model--the probability mixture. There are many clustering algorithms [Hartigan], [Patrick], [Duda], each having specific characteristics and a performance that greatly depends on the character of the data. Almost all of the algorithms require a specification of the number of clusters, and frequently other parameters as initial values for the iteration.

The results of clustering for a given data set strongly depend on the specified number of clusters and on a choice of the algorithm and its initial parameters. To be able to apply clustering in a meaningful way to a set of data with little known characteristics it is necessary to have a method of evaluating and comparing different results and a criterion for selecting one of them as the best. Criteria related to likelihood or quadratic error function, such as the minimum of the average Euclidian distance within clusters cannot be used to determine the number of clusters [Friedman].

Other techniques, based on a heuristic application of the ratio of the between- to within-cluster scatter measures [Fukanaga], [Coleman], or a more rigorous method of Vogel and Wong [Vogel] based on pseudo F-statistic can help to determine the number of clusters, but their application is limited to the cases where either Euclidian or Bhattacharyya distance is used and the clusters are oval-shaped.

The minimal representation criterion provides a natural and the most general method for selecting a cluster configuration whenever the clustering is used in a probabilistic context. The application of the minimal representation criterion to clustering problems has been proposed by Segen and Sanderson [Minimum], and a related approach has been used by Wallace and Boulton [Boulton].

For a given set of k clusters the probability of x can be written as:

$$P_q(x) = \sum_{i=1}^k P_q(x|i) \cdot P_q(i)$$

where $P_q(x|i)$ is the probability, given that x is a member of the cluster i , and $P_q(i)$ is the probability of the cluster i .

Let $a_q(x)$ be a decision function which assigns x to one of the clusters. If the form of the $P_q(x|i)$ is such that $P_q(x|i) = 0$ if $i \neq a_q(x)$, then

$$P_q(x) = P_q(x|a_q(x)) \cdot P_q(a_q(x)),$$

and since the observations x_i are assumed to be independent, then

$$\log P_q(x^n) = \sum_{i=1}^n [\log P_q(x_i|a_q(x_i)) + \log P_q(a_q(x_i))].$$

The set of parameters q contains the information about $P(x|i)$, $P(i)$, and the number of clusters; its specific form and the value of $S(q)$ depends on the assumed general form of $P(x|i)$ and $P(i)$. The cluster configuration is given by the decision function $a(x)$. To select the best configuration of clusters we minimize $r(q, x^n)$ for each given configuration and take the one giving the global minimum. This method can be used to discriminate among results of different clustering algorithms and different assumptions about the number of clusters and the initial parameters. The clustering in this case can be interpreted as a search-space decomposition in the problem of searching for the best model. The following simple example illustrates the use of the above technique.

Example 1

Assume that we have 100 observations--natural numbers, distributed according to the histogram in Figure CL1. We want to decide whether to represent the observation as one cluster (0-300), two clusters (0-100) and (200-300), three clusters (0-50), (50-100), (200-300) or four clusters (0-50), (50-100), (200-250) and (250-300). The distribution within the clusters is assumed to be uniform. The set q consists of the following parameters:

k - number of clusters

B_i - lower end of the cluster i

T_i - width of the cluster i

$B_i + T_i < B_{i+1}$

$P(1), P(2), \dots, P(k-1)$ - cluster probabilities

The decision function $a(x)$ represents the evaluated cluster assignment.

The probability $P(x|i)$ is

$$P_q(x|i) = \begin{cases} 1/T_i & \text{if } a(x) = 1 \\ 0 & \text{otherwise} \end{cases}$$

So,

$$\begin{aligned} -\log P_q(x^n) &= \sum_{i=1}^k \log(1/T_i) \cdot \sum_{i=1}^k n_i \log P(i) \\ &= \sum_{i=1}^k n_i \log(T_i/P_i) \end{aligned}$$

and

$$s(q) = s(k) + \sum_{i=1}^k [s(B_i) + s(T_i)] + \sum_{i=1}^{k-1} s(P_i)$$

Minimizing $r(q, x^n)$ for each of the compared cases will give:

1 cluster:

$$B_1 = 0, T_1 = 300, P(1) = 1$$

$$-\log P_q(x^n) = 100 \log 300 = 822$$

$$S(q) = S(1) + S(0) + S(300) = 22$$

$$r(q, x^n) = S(q) \cdot \log P_q(x^n) = 844$$

2 clusters

$$B_1 = 0, T_1 = 100, B_2 = 200, T_2 = 100, P(1) = P(2) = 1/2$$

$$-\log P_q(x^n) = 100 \log 200 = 764$$

$$S(q) = S(2) + S(0) + S(200) + 2S(100) + S(0.5) = 54$$

$$r(q, x^n) = 818$$

and similarly for 3 clusters $r(q, x^n) = 848$, and for 4 clusters $r(q, x^n) = 880$.

These results are plotted in Figure CL2 where the optimal two cluster case is clearly indicated by a minimum.

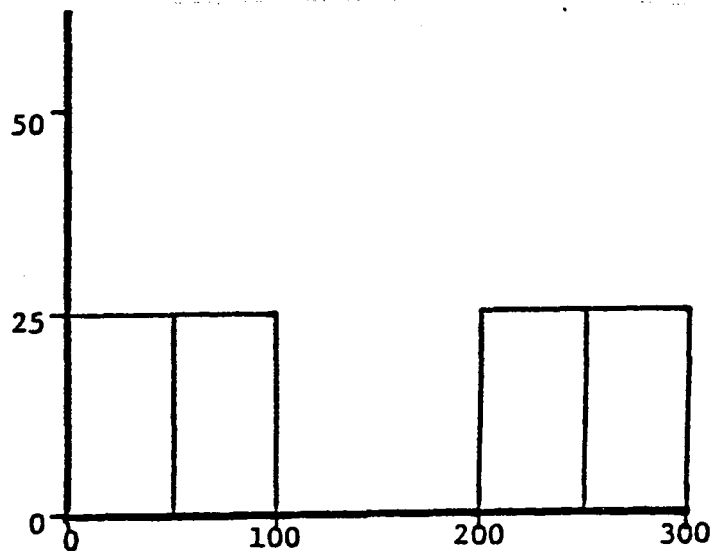


Figure 4: Histogram of observations for Example 1.

Example 2

100 two-dimensional values were generated by a mixture of 3 Gaussian distributions. Each of the component distributions had independent marginals with equal variance. The means and the variances of the component distributions were:

1. Mean (0, 0), variance 1
2. Mean (10, 0), variance 9
3. Mean (0, 5), variance 4

Each of the components had equal probability of 0.333... . The generated values are shown in Figure CL3. The k-means [Hartigan] clustering algorithm was applied to 30, 40, 60, and 100 values using

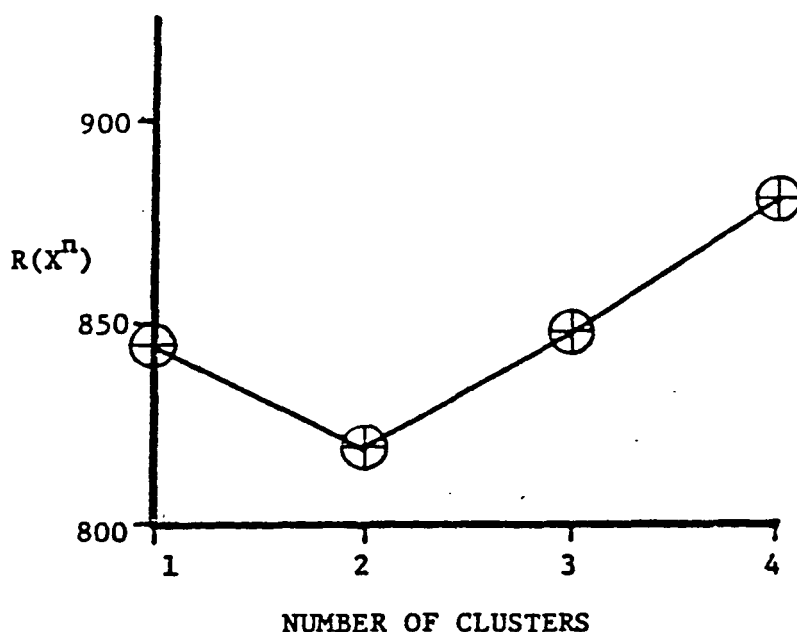


Figure 5: Representation size function vs. the number of clusters in Example 1.

$k = 1, 2, \dots, 10$. Each cluster configuration was evaluated by an algorithm computing $r(q, x_n)$ and the results are plotted in Figure CL4 using the logarithmic transformation to show clearly the minimum. The numbers above plots show how many observations were used. The minimum in each case was reached for 3 clusters, so the program discovered the correct number of the mixture components.

4. Complex Markov chain

A complex Markov chain model of order- k for a sequence of symbols from a finite alphabet has a form:

$$P(x_n | x^{n-1}) = P(x_n | x_{n-1}, x_{n-2}, \dots, x_{n-k})$$

The maximum likelihood procedure can be used to estimate the transition probabilities for a given order. If the order is unknown and the maximum likelihood method is used to determine it, as a result one will almost always obtain the highest possible order, which makes it not very meaningful. The minimal representation method can provide a simple answer to this problem.

5. Multiple shifts in the mean value

Assume that n observations are independently distributed and $x_i \propto N(m_i, \sigma^2)$ where $m_i = h_j$ for $t_j \leq i < t_{j+1}$, $j = 0, 1, \dots, k$ and $1 = t_0 < t_1 < \dots < t_{k+1} = n+1$, i.e., the mean shifts k times during the observation period, and k can be $0, 1, \dots, n-1$. If the number of shifts is unknown, the application of maximum likelihood method to identify this model would always give a meaningless result $k = n-1$. The minimal representation method provides an approach to this problem, although additional techniques may have to be used to reduce the search.

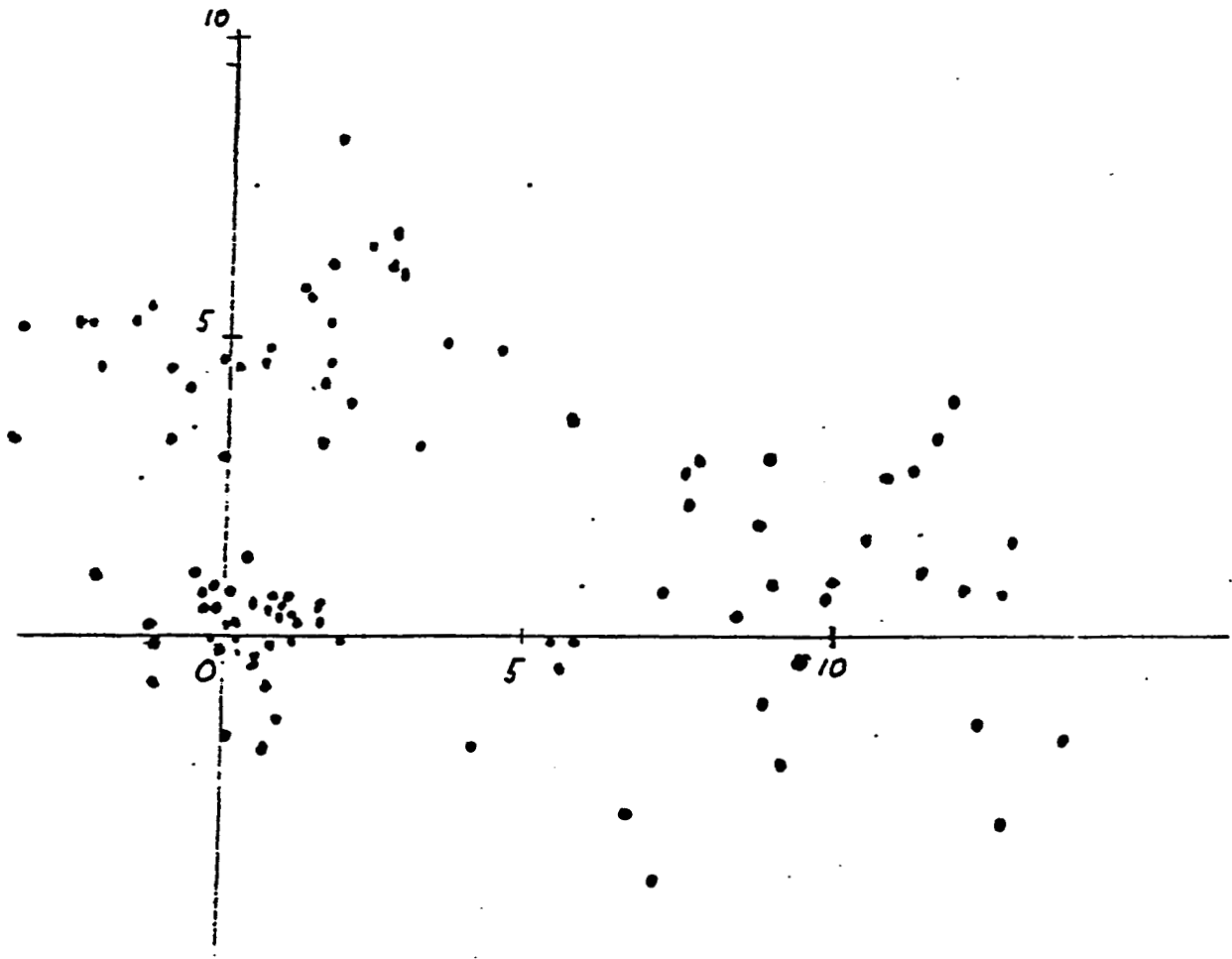


Figure 6: Data for clustering in Example 2

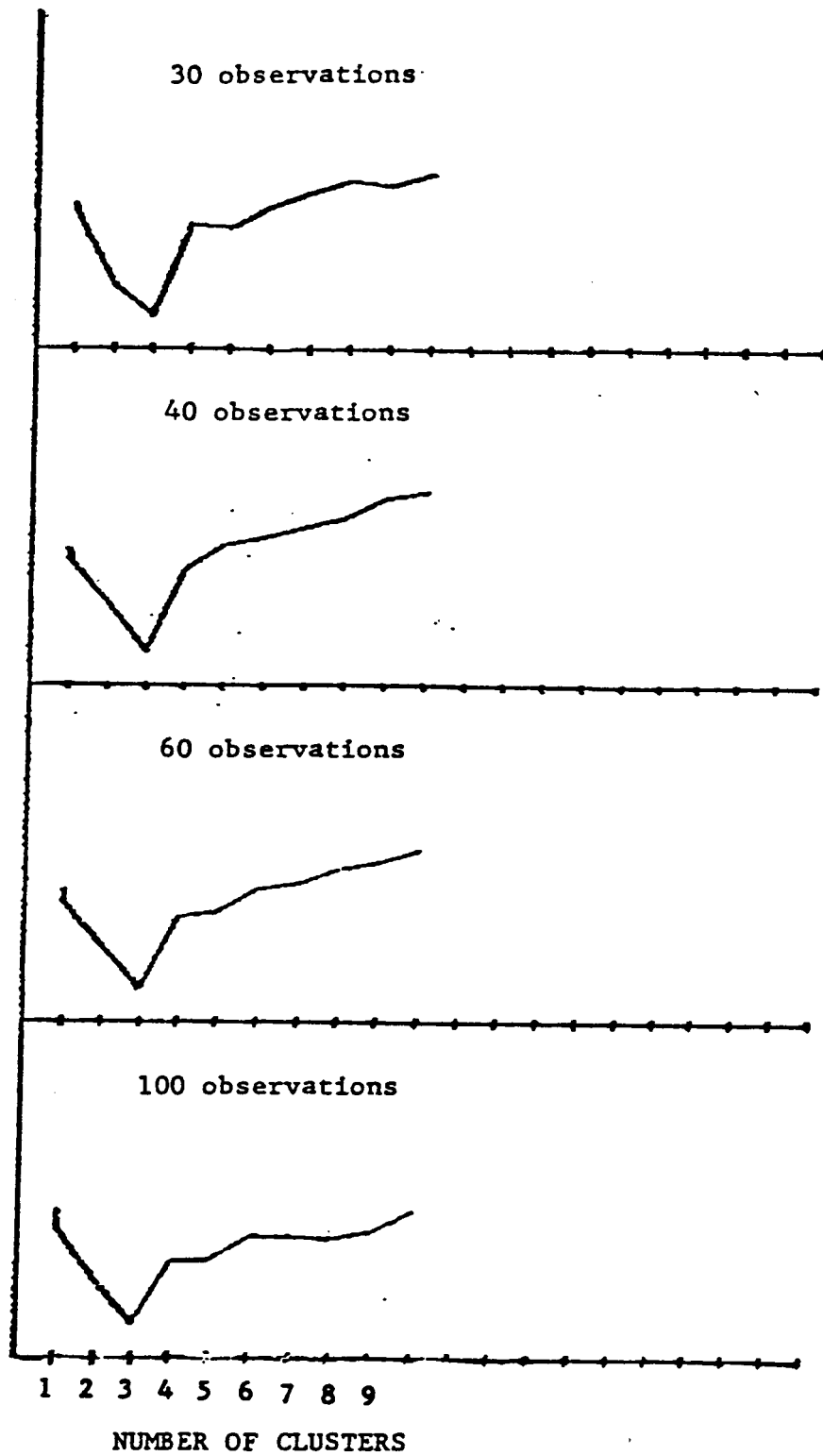


Figure 7: Representation size evaluation of clustering, Example 2

A large body of applications may be found among problems of concept formation, inductive inference or learning which are of strong interest to artificial intelligence. These problems are presently approached on a more heuristic level [Lenat], [Buch2] [Quinlan].

- [1] H. Akaike.
A new look at the statistical model identification.
IEEE Trans. Autom. Contr. AC-19:716-723, 1974.
- [2] T.W. Anderson.
The statistical analysis of time series.
John Wiley, 1971.
- [3] C.S. Wallace and D.M. Boulton.
An informaton measure for classification.
Comp. Journal 11(2):185, 1968.
- [4] G.E.P. Box and G.M. Jenkins.
Time series analysis, forecasting and control.
Holden-Day, San Francisco, 1970.
- [5] B.G. Buchanan and T.M. Mitchel.
Model directed learning of production rules.
In Waterman and Hayes-Roth (editor), *Pattern directed inference systems*, . Academic Press,
New York, 1978.
- [6] G.J. Chaitin.
A theory of program size formally identical to information theory.
J. Ass. Comp. Mach. 22(3):329-340, 1975.
- [7] G.J. Chaitin.
Information-theoretic computational complexity.
IEEE Trans. Inform. Theory IT-20:10-15, 1974.
- [8] G.B. Coleman and H.C. Andrews.
Image segmentation by clustering.
Proc. IEEE 67(5):773-785, 1979.
- [9] R.O. Duda and P.E. Hart.
Pattern classification and scene analysis.
John Wiley, New York, 1973.
- [10] H.P. Friedman and J. Rubin.
On some invariant criteria for grouping data.
J. Amer. Stat. Assoc. 62:1152-1178, 1967.
- [11] K. Fukanaga.
Introduction to statistical pattern recognition.
Academic Press, New York, 1972.
- [12] J.A. Hartigan.
Clustering algorithms.
John Wiley, New York, 1975.

- [13] R.L. Kashyap and A.R. Rao.
Dynamic stochastic models from empirical data.
Academic Press, New York, 1976.
- [14] A.N. Kolmogorov.
On the logical basis of information theory and probability theory.
IEEE Trans. Inform. Theory IT-14:662-664, 1968.
- [15] T.S. Kuhn.
The structure of scientific revolutions.
University of Chicago Press, Chicago, 1962.
- [16] S. Kullback.
Information theory and statistics.
John Wiley, New York, 1959.
- [17] D.B. Lenat and G. Harris.
Designing a rule system that searches for scientific discoveries.
Technical Report, Carnegie-Mellon Univ. Comp. Sci. Dept., April, 1977.
- [18] J. Segen and A.C. Sanderson.
A minimal representation criterion for clustering.
In *Computer Science and Statistics: 12 Annual Symposium on the Interface.* University of Waterloo, Waterloo, May, 1979.
- [19] E.A. Patrick.
Fundamentals of pattern recognition.
Prentice-Hall, Englewood Cliffs, N.J., 1972.
- [20] J.R. Quinlan.
Induction over large data bases.
Technical Report STAN-CS-79-739, Stanford Univ. Computer Sci. Dept., 1979.
- [21] J. Rissanen.
Modeling by shortest data description.
Automatica 14:465-471, 1978.
- [22] J. Rissanen.
Consistent order estimates of autoregressive process by shortest description of data.
Unpublished report. The author is at IBM San Jose research center.
- [23] J. Rissanen.
Shortest data description and consistency of order estimates in ARMA-processes.
In *Intern. Symp. on Optimiz. and Stoch. Processes.* IRIA, Paris, 1978.
- [24] G. Schwarz.
Estimating the dimension of a model.
Ann. Stat. 6(2):461-464, 1978.

- [25] R. Shibata.
Selection of the order of an autoregressive model by Akaike's information criterion.
Biometrika 63(1):117-126, 1976.
- [26] R.J. Solomonoff.
A formal theory of inductive inference.
Inform. and Control 7:1-22 & 224-254, 1964.
- [27] R.J. Solomonoff.
Complexity-based induction systems: Comparison and convergence theorems.
IEEE Trans. Inform. Theory IT-24:422-432, 1978.
- [28] G. Toussaint.
Bibliography on estimation of misclassification.
IEEE Trans. Inform. Theory IT-20:472-479, 1974.
- [29] M.A. Vogel and A.K.C. Wong.
PFS clustering method.
IEEE Trans. Pattern Anal. Mach. Intel. PAMI-1:237-245, 1979.
- [30] D.G. Willis.
Computational complexity and probability constructions.
J. Ass. Comp. Mach. 17(2):241-259, 1970.