Estimating the Markov Order of Symbolic Sequences Using Surrogate Data

Marcel J. van der Heyden [†] [*] Cees Diks [†]
Bart P.T. Hoekstra [†] Jacob DeGoede [†]

9th January 1997

Abstract

Symbolic sequences can be statistically characterized by their m^{th} order Markov structure, which is defined in terms of m^{th} order transition probabilities , as well as derived quantities such as the m^{th} order (conditional) entropy. A methodology for testing the null hypothesis that a symbolic sequence has m^{th} order Markov structure is described which utilizes the bootstrap resampling method and is applied to binary sequences with known order. We determine the size and power of the symbolic surrogate data test, as well as that of the χ^2 test for Markov order. The effects of conditioning on nuisance parameters and the sufficient statistic in this problem of statistical inference are also investigated. We conclude that conditioning on the sufficient statistics increases the power of the symbolic surrogate test and that it outperforms the χ^2 test, especially for short sequences.

[*] From the Graduate School Neurosciences Amsterdam

[†] Dept. Physiology, Leiden University, PO box 9604, 2300 RC Leiden, the Netherlands
Tel: + 31 (0)71 5276751, Fax: + 31 (0)71 5276782
WWW: `www.leidenuniv.nl/medfac/fff/groepc/chaos/`
Email: `heijden/bhoekstra/diks/degoede@rullf2.medfac.leidenuniv.nl`

Contents

1 Introduction **3**

2 Markov Chains and Entropy **4**

3 Testing for Markov Order **6**

3.1 Entropy Test 6

3.2 χ^2 Test 7

4 Symbolic Surrogate Test **7**

4.1 Symbolic Surrogates 8

4.1.1 Typical Realization Surrogates 9

4.1.2 Constrained Probabilities 9

4.1.3 Constrained Realizations 10

5 Application to Test Data **11**

5.1 Test Data 11

5.2 Entropy 11

5.3 χ^2 Test 12

5.4 Symbolic Surrogate Data Test 13

5.4.1 Typical Realization Surrogates 13

5.4.2 Constrained Probabilities 14

5.4.3 Constrained Realizations 15

6 Discussion **16**

1 Introduction

The analysis of a symbolic representation of a dynamical system or experimental data set may have several advantages. Some systems such as DNA, languages, spin glasses and ion channel dynamics are inherently discrete. Other systems are continuous but their properties or behavior suggest that a symbolic representation may be useful. Examples are systems that intermittently switch from one volume in phase space to another distant one or that poses symmetries in their structure or behavior, suggesting that certain features of such system may be well represented using a symbolic representation. Furthermore, the use of symbolic sequences in the analysis of noisy systems has been advocated by Crutchfield and Packard [7] and Tang *et al.* [20]. Finally, a coarse graining of the phase space yielding a symbolic representation may improve the statistics in certain estimation or inference problems.

When the phase space of a time continuous dynamical system is partitioned using a *generating partition* it can be shown that certain properties of the original system, like the entropy, are invariant under the transformation to a statistical symbolic description [9]. In general the selection of a partition of the state space of the system or a given (experimental) data set is far from trivial, in the sense that the properties of the symbolic system or data set depend on the partition chosen and may thus have little meaning by themselves. However, in certain cases one can select partitions on the basis of physi(ologi)cal arguments that highlight some features of a system or data set while effectively assuming that other structure in the system or data are due to, in a particular context, uninteresting fluctuations.

The description of a system generating a symbolic sequence in terms of a Markov process is a very general one and can elucidate many properties of the system underlying the data. Much attention has been paid to the correlation properties of symbolic sequences such as DNA [13, 16], natural languages [8] and neural spike trains [4] and an estimate of the order of the Markov chain observed may provide important additional information. Here we propose a method which tests the null-hypothesis (H_0) of the data being of n^{th} Markov order and uses a bootstrapping technique [10] to assign significance levels to the results obtained.

After summarizing some properties of Markov chains and their entropies, two existing tests for Markov order, used for comparison with the symbolic surrogate method, will be briefly described. The symbolic surrogate test will be described in more detail and the role of conditioning on nuisance parameters or the sufficient statistic will be discussed with reference to recent statements made in the context of testing nonlinear structure in continuous time series. After application of the tests to Markov sequences with known order and calculation of the power and size of the tests we discuss these results and suggest possible practical applications in which the use of symbolic surrogates may prove advantageous.

2 Markov Chains and Entropy

A symbolic sequence

$$x_i \qquad i = 1, \cdots, N \tag{1}$$

of which the elements are drawn from an alphabet consisting of λ members A_j

$$A_1, \cdots, A_\lambda. \tag{2}$$

can be characterized statistically by its Markov structure. We will call a sequence composed of independent identically distributed (iid) elements to be of 0^{th} Markov order. One can further define n^{th} order ($n \geq 1$) Markovian sequences, which have the property that the probability $p(x_i)$ is only dependent on its past values $x_{i-n}, \cdots, x_{i-1}$:

$$p(x_i) \quad = \quad f(x_{i-n}, \cdots, x_{i-1}). \tag{3}$$

Such n^{th} order Markov chains are defined through the n^{th} order conditional probabilities (also called transition probabilities)

$$p(x_i = A_r | x_{i-n} = A_s, \cdots, x_{i-1} = A_t) \tag{4}$$

which will in the remainder be denoted using the following notation

$$p(a_i | a_{i-n}, \cdots, a_{i-1}). \tag{5}$$

The conditional probabilities of an n^{th} order Markov chain have the following property:

$$p(a_i | a_{i-m}, \cdots, a_{i-n}, \cdots, a_{i-1}) \quad = \quad p(a_i | a_{i-n}, \cdots, a_{i-1}) \tag{6}$$

for $m \geq n$.

Another way to characterize (symbolic) sequences is by their m^{th} *order entropies* which are defined as

$$H_m \quad = \quad - \sum_{a_{i-m}, \cdots, a_{i-1}} p(a_{i-m}, \cdots, a_{i-1}) \log p(a_{i-m}, \cdots, a_{i-1}). \tag{7}$$

where the summation is over all combinations $a_{i-m}, \cdots, a_{i-1}$ for which $p(a_{i-m}, \cdots, a_{i-1}) \neq 0$. These entropies may be interpreted as the mean uncertainty in the prediction of sub-sequences of length m. Another measure which represents the mean uncertainty in the prediction of x_i when $x_{i-m}, \cdots, x_{i-1}$ are known is the m^{th} *order conditional entropy* which is defined as

$$h_m \quad = \quad H_{m+1} - H_m \tag{8}$$

and we define $h_0 = H_1$. It can be expressed in terms of the m^{th} order conditional probabilities as follows:

$$\begin{aligned} h_m &= - \sum_{a_{i-m},\cdots,a_{i-1}} p(a_{i-m},\cdots,a_{i-1}) \times \\ &\quad \sum_{a_i} p(a_i|a_{i-m},\cdots,a_{i-1}) \log p(a_i|a_{i-m},\cdots,a_{i-1}). \end{aligned} \tag{9}$$

The metric entropy is defined by the limit

$$h = \lim_{m\to\infty} \frac{H_m}{m} = \lim_{m\to\infty} h_m. \tag{10}$$

It was shown by Shannon [19] that $h_{m+1} \leq h_m$ for all m. For n^{th} order Markov chains the following applies to the conditional entropies:

$$h_m = h_n \tag{11}$$

for $m \geq n$.

In practice one estimates the m^{th} order entropy and the m^{th} order conditional entropy using the sample probabilities $\widehat{p}$ which are estimated on the basis of the *transition counts*

$$n_{a_{i-m},\cdots,a_i} \tag{12}$$

which denote the number of (in this study overlapping) tuples $a_{i-m},\cdots,a_i$ found in the data set. For an n^{th} order Markov chain the $(n+1)^{st}$ order transition counts $n_{a_{i-n},\cdots,a_i}$ together with the first n elements (the initial conditions $x_1,\cdots,x_n$) are a *sufficient statistic* that completely characterizes such chain statistically.

The m^{th} order sample joint probabilities and m^{th} order sample conditional probabilities are estimated using their maximum likelihood estimators

$$\widehat{p}(a_{i-m},\cdots,a_{i-1}) = \frac{n_{a_{i-m},\cdots,,a_{i-1}}}{N_m} \tag{13}$$

$$\widehat{p}(a_i|a_{i-m},\cdots,a_{i-1}) = \frac{n_{a_{i-m},\cdots,,a_{i-1},a_i}}{n_{a_{i-m},\cdots,a_{i-1}}} \tag{14}$$

where N_m is the total number $(N-m+1)$ of m-tuples that can be extracted from a sequence of length N.

In this study the probabilities $\widehat{p}$ are estimated from the data set being made circular (period N, $N_m = N$) so that varying number of tuples of different lengths that can be extracted from the data do not have to be taken into account. Thus, notational and computational complexity is reduced and possible end-effects are avoided. In particular, it is thus ensured that no absorbing states are introduced into the transition probability matrix. All states communicate and thus belong to the same equivalence class which means that the chain is by definition irreducible and ergodic. Note that the application of circular data does not alter the results presented here in any significant way, but that it may have a larger effect when there are strong autocorrelations present in a relatively small data set.

3 Testing for Markov Order

In some cases the goal may not be to characterize a symbolic sequence by accurately estimating the (conditional) probabilities for explicit model construction. Rather one may want to merely asses the Markov order of the symbolic data set. This may be the case prior to the construction of a model and estimation of its parameters.

Various methods for testing for Markov structure have been proposed such as χ^2 distributed likelihood ratio [12] and goodness-of-fit tests [2, 3] which can be shown to be asymptotically equivalent [1]. These methods test the H_0 that the observed sequence is a realization of a Markov process of given order against the alternative hypothesis (H_1) that it is of higher order. The use of information criteria is well known in the context of order estimation for continuous time series and has also been applied to symbolic sequences [14, 23]. Also tests have been described which test the H_0 of a Markov process with known transition probabilities, using Ψ^2 test statistics [5]. Finally, information theoretical quantities like entropy can also provide the means for testing for Markov order when taking into consideration the properties of the conditional entropy for Markov sequences. It can even be shown that such methods are related to likelihood ratio test [5].

A review and discussion of all methods mentioned above is beyond the scope of this paper. However, we will discuss and apply the behavior of the conditional entropy h_m as a function of the order m as well as the χ^2 test to create a context for the symbolic surrogate methods for testing the Markov order.

3.1 Entropy Test

The properties of the conditional entropy h_m (Eqn. 11) provide a way to test Markov order of a given sequence by looking at its behavior for increasing m. If h_m converges to a fixed value for m larger than some n one can conclude that the sequence is of Markov order n.

In practice, however, this may not be a trivial task. The estimation of the higher order entropies requires the estimation of higher order (conditional) probabilities which can be difficult to do accurately when a limited amount of data is available. Too little data can result in a large underestimation the entropy so that spurious convergence may be observed [11]. Various finite sample corrections for sequence analysis have been proposed [11, 15], but often assumptions are made which may not always be reasonable or useful (see however [15]). Furthermore, little is known on the distributions of such entropy estimations making the estimation of significance levels difficult.

3.2 χ^2 Test

Traditionally, a lot of emphasis has been placed on the application of χ^2 tests for statistical inference about Markov chains (see e.g. [1, 2, 3, 5]). Such tests can be used to test the H_0 of the data being of m^{th} Markov order against the H_1 of Markov order $m+1$. In that case the statistic is given by

$$\begin{aligned}\sum_{a_{i-m-1},\cdots,a_i} \quad & \frac{[N\widehat{p}(a_{i-m-1},\cdots,a_i)}{N\widehat{p}(a_{i-m-1},\cdots,a_{i-1})\widehat{p}(a_i|a_{i-m},\cdots,a_{i-1})} - \\ & \frac{N\widehat{p}(a_{i-m-1},\cdots,a_{i-1})\widehat{p}(a_i|a_{i-m},\cdots,a_{i-1})]^2}{N\widehat{p}(a_{i-m-1},\cdots,a_{i-1})\widehat{p}(a_i|a_{i-m},\cdots,a_{i-1})} \qquad (15)\end{aligned}$$

which is asymptotically χ^2 distributed with $d_{m+1}-d_m$ degrees of freedom where

$$\begin{aligned} d_{m+1} &= \lambda^{m+2} - \lambda^{m+1} \\ d_m &= \lambda^{m+1} - \lambda^m . \end{aligned}$$

The summation in Eqn. 15 is performed over the indices $a_{i-m-1},\cdots,a_i$ for which the denominator is > 0. When the sample probabilities

$$\begin{aligned} \widehat{p}(a_i|a_{i-m},\cdots,a_{i-1}) &= 0 \\ \widehat{p}(a_{i-m-1},\cdots,a_i) &= 0 \end{aligned}$$

then d_m and d_{m+1} respectively are decremented by 1.

When applying this test to a data set of unknown order the procedure consists in starting with $m=0$. When the H_0 of 0^{th} order can be rejected, m is increased by 1 and the test is repeated. The estimated order is then chosen as the smallest value of m for which the H_0 of m^{th} order can not be rejected. Significance levels for the χ^2 test are obtained using the incomplete Gamma function [17].

4 Symbolic Surrogate Test

In this section a novel way to test the Markov order of a symbolic sequence using a bootstrapping method [10] is proposed which is in some respects analogous to the method of surrogate data [21] for testing continuous data sets for nonlinearity.

In general, the surrogate data procedure is to specify a H_0 which is a class of models which is sought to reject as a probable model for the data. A number of realizations of the H_0 process(es) is generated - the surrogate data and a test statistic S is selected. Its value for the original data S^d as well as for the ensemble of surrogate data $\{S_i^s\}_{i=1}^B$ is estimated. It can be assessed at which significance level the original data are likely not to be a realization of a process in the H_0 class, for instance by determining the rank of S^d with respect to the elements of the set $\{S_i^s\}_{i=1}^z$.

An important property of tests is the *power* of the test; the fraction of trials that the null is correctly rejected. The power of a test is a function of S and the properties of the data that are tested but also the method to generate the surrogate data can have a large influence on the power. The fraction of times that the H_0 is rejected when the data actually comply to the H_0 is called the *size* of the test. Caveats and subtleties of the generation of surrogates are discussed by Theiler and Prichard [22] and has analogues to the case of Markov order estimation of symbolic data discussed here.

In the process of statistical inference it is often attempted, implicitly or explicitly, to eliminate the influence of nuisance parameters by conditioning on the sufficient statistic so as to enable the computation of accurate approximations to densities [18]. As a result of conditioning one may construct tests that have increased power with respect to those that do not condition and which have a true size that better approximates the nominal value. An example of such a case was presented by Prichard and Theiler [22] in the context of testing for nonlinearity in continuous data sets. It will be shown that in the problem of testing the Markov order of a symbolic sequence presented here conditioning has very similar effects. The use of the term "constrained" has been adapted from [22] meaning roughly the conditioning on nuisance parameters or sufficient statistics.

4.1 Symbolic Surrogates

The procedure to test the H_0 that the data are of n^{th} Markov order consists of generating an ensemble of realizations with n^{th} order Markov properties matching those of the data set to be tested. These properties are the n^{th} order transition probabilities in Eqn. 5. Since the true transition probabilities are unknown, the maximum likelihood sample estimates (see Eqn. 13) of the n^{th} order transition probabilities from the original data set are used instead to generate our symbolic surrogate data.

We describe three ways to generate surrogate data, or surrogate conditional probabilities, with known Markov order called *typical realizations*, *constrained probabilities* and *constrained realizations*.

- **Typical n^{th} order realization** surrogate data are obtained by fitting the "best" n^{th} order model (i.e. the maximum-likelihood sample conditional probabilities) to the data and then use this model to generate surrogate realizations using as the first n elements of our surrogates the first n elements of our original data set. Note that in general the transition probabilities estimated from the surrogates are different from those estimated from the original data due to random fluctuations.
- **The constrained probabilities** method yields transition probabilities of order $n + 1$ which have *identical* n^{th} order sample properties as the original data. More specifically, the n^{th} order sample transition probabilities estimated from the surrogate $m+2$-tuples (by using

either the first or the last $n+1$ symbols in those tuples) are identical to those estimated from the original data set. However, there is no unique way, in general, to construct a symbolic sequence that corresponds to this set of probabilities.

- **The constrained realization** method *does* yield a sequence with identical n^{th} Markov order properties as the original data but may be computationally intensive to construct since there is no one-shot method to constructing the symbolic surrogate sequence. In the worst case (small data sets with long autocorrelations) there may not even be enough surrogate realizations to yield test results at an a-priori specified significance level.

Having constructed the surrogate realizations or transition probabilities a test quantity is selected which is dependent on (sensitive to) Markov properties of higher order than the surrogates and can thus be used to detect differences between S^d and the ensemble of S^s. In this study the $(n+1)^{st}$ order conditional probability h_{n+1} as defined in Eqn. 8 is used as a the test statistic.

After the calculation of h^d_{n+1} of the data set and B of its surrogates we rank-order the B h^s_{n+1}'s and determine the rank r of the h^d_{n+1}. It is ranked 1 if it is the smallest, and $B+1$ if it is the largest in the whole set. Since it can be shown that asymptotically h^d_{n+1} is always $\leq h^s_{n+1}$ we can use a one-sided test and reject the H_0 with a significance level of $\frac{r}{B+1}$.

4.1.1 Typical Realization Surrogates

We generate a typical n^{th} order surrogate by taking the first n elements of the surrogate realization equal to the first n elements of the original sequence. Subsequent elements of the surrogate realization are then generated using a random number generator and the sample n^{th} order transition probabilities. The result is a symbolic surrogate sequence of at most n^{th} order from which the transition probabilities can be estimated which are used in the calculation of S^s.

Note however, that even when the data actually comply to the H_0 the n^{th} order sample probabilities from the surrogates may differ from those of the original series, in particular when they are small. This may introduce a relatively large variance in S^s which may have an adverse effect on the power of this test.

4.1.2 Constrained Probabilities

The method of constrained probabilities does not yield an n^{th} Markov order surrogate sequence from which the surrogate transition probabilities of order $n+1$ can be estimated. Instead it directly results in the transition probabilities of order $n+1$ but with the exact n^{th} order sample statistical properties of the original data set. This method does not however condition on the complete sufficient statistics under the H_0 since it does not use the information contained in the first n elements of the original sequence. There may be realizations corresponding to this set of transition probabilities that do not have the same initial conditions as the original data or there may not even be a realization that would yield the surrogate sample transition probabilities.

From the original (circular) data set all tuples $(a_i, \cdots, a_{i+n+2})$ of length $n+2$ are extracted. These tuples contain the information of markov order $n+1$. With this set of tuples one can estimate all sample probabilities of order smaller than $n+1$ by simply selecting subsets from the tuples. For instance, the 0^{th} order probabilities can be constructed from the set of the i^{th} elements in every tuple, and the 1^{st} order transition probabilities can be constructed from the set of every $(i, i+1)$ doublet in every tuple.

We now construct a surrogate set of tuples by first splitting up the set of $n+2$-tuples $(a_i, \cdots, a_{i+n+2})$ into two new sets of $n+1$-tuples: $(a_i, \cdots, a_{i+n+1})$ and $(a_{i+1}, \cdots, a_{i+n+2})$. Next, the surrogate set of $n+2$-tuples is constructed by taking a $n+1$-tuple $(a_i, \cdots, a_{i+n+1})$ from the first set of $n+1$ tuples (without replacement) and randomly drawing (without replacement) a tuple $(a_{i+1}, \cdots, a_{i+n+2})$ which' elements $a_{i+1}, \cdots, a_{i+n+1}$ exactly match the corresponding ones of the first tuple. That is, the last n elements in the first $n+1$-tuple are identical to the first n in the second $n+1$-tuple. The new set of $n+2$-tuples has by construction at most Markov order n and yields identical sample properties of order n and lower.

4.1.3 Constrained Realizations

It is possible to construct a surrogate symbolic *sequence* of a Markov order of at most n by extracting the set of $n+1$ tuples from the original sequence. Then a surrogate sequence is constructed by starting with an $n+1$-tuple of which the first n elements are identical to those of the original time series. Subsequently, $n+1$- tuples are randomly drawn (without replacement) of which the first n elements correspond to the last n in the surrogate sequence until all $n+1$ tuples are used (thus having constructed a surrogate realization) or until no matching $n+1$ tuples are available anymore in which case a valid realization was not found. These surrogate realizations have by construction a Markov order of not larger than n and identical n^{th} and lower order sample properties.

Note that finding a realization can take up a lot of computer time, especially when n is large and the data are highly correlated. Also, when the data are highly correlated it is conceivable that there are only very few realizations possible. The same holds for the constrained probabilities, but to a lesser extend.

5 Application to Test Data

In this section we apply the tests described above to binary sequences generated by well defined mathematical models generating sequences with known Markov properties.

5.1 Test Data

A straightforward way to create (binary) data with known Markov order is by assigning a random number q to the nth order conditional probabilities in the following way:

$$\begin{aligned} p(a_i|a_{i-n},\cdots,a_{i-1}) &= q \\ p(\bar{a}_i|a_{i-n},\cdots,a_{i-1}) &= 1-q \end{aligned}$$

for all combinations $a_{i-1},\cdots,a_{i-n}$. Here the value of q is in principle arbitrary but in this study different models were generated by randomly drawing q from a uniform distribution $[0.16, 0.84]$. Values for q too close to 0 or 1 were not used in order to avoid generating test data with strong autocorrelations. Furthermore, the choice of random number generator can have a large influence on the results, particularly when a bad one is chosen with a short period or relatively strong correlation between consecutive numbers. The random number generator provided with the compiler was used in this study and it was checked that the results were not significantly different when using the `ran3` routine described in [17] in a number of representative cases.

From the set of conditional probabilities test data of Markov order n were generated by selecting randomly the first n symbols in the sequence and using the conditional probabilities to generate subsequent symbols in the sequence. The Markov chain was allowed to converge to its stationary distribution by discarding the first 4096 samples as a transient.

For all results described below 1000 symbolic sequences with known Markov order of length 512 coming from a 1000 different, randomly selected, models were generated. Thus, in a sense the power and size calculated below are *averages* over a class of Markov models of fixed order. Identical models and sequences of a given order were used for all results described below to allow for an optimal comparison of the different methods.

5.2 Entropy

Since for n^{th} order Markov chains $h_m = h_n$ for $m \geq n$, a plot of h_m as a function of m can be used to estimate the Markov order by visual inspection. It was tested whether this convergence can indeed be observed for the test data. The results for 2^{nd} and 5^{th} order binary test data are displayed in Figs. 1. The data points plotted are averages over realizations of the

1000 random models and the error bars indicate the standard deviation of these values.

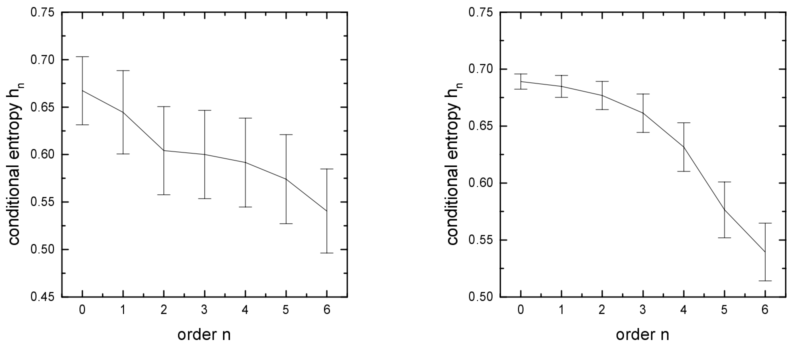

Figure 1: *Plots of the conditional entropy h_m against Markov order m for binary test data of Markov order 2 (left) and 5 (right). Plotted are the mean values over realizations of length 512 of 1000 different models with random transition probabilities. The error bars denote the standard deviations of the estimated entropies. We did not apply any methods to correct for possible finite sample effects that may bias the h_n estimates.*

For the 2^{nd} order sequences we see a change in the slope of the figure at the correct value of n, but also a slow continuing decrease of h_n with increasing order n. For the 5^{th} order models convergence of the values of h_n is invisible and it is not possible to make an estimate of the Markov order on the basis of that figure. Instead a continuing decrease of h_m is observed due to finite sample effects. Using longer sequences the convergence of h_m became clearer.

The application of h_n in the estimation of the Markov order becomes even more difficult when the number of elements λ in the alphabet becomes larger or the Markov order of an (experimental) is unknown a priori. This, and the fact that one cannot make statistical inference on the basis of the values of h_n since its standard deviation is unknown, makes the use of h_n for estimating the Markov order of a given data set difficult.

5.3 χ^2 Test

For the χ^2 test for Markov order the test data were used to assess the power and size of this test for comparison with the symbolic surrogate data test. We rejected the H_0 at a significance level of 0.05. The power of the χ^2 test was determined by testing the H_0 of n^{th} Markov order against the H_1 of Markov order $n+1$ using sequences of order $n+1$, where n was varied from 0 to 5. The size of the test was determined by testing the H_0 of n^{th}

Markov order against the H_1 of Markov order $n+1$ using sequences of order n, where n was again varied from 0 to 5.

The results are displayed in Table 1. The top row indicates the Markov order of the H_0 that is tested. The numbers in the bottom rows denote the power and size of the χ^2 test. The test results are displayed with 2 digit accuracy since the standard deviation of the power and size is approximately $\sqrt{\frac{1}{1000}} = 0.03$.

Table 1: *Results of the χ^2 test for Markov order applied to the binary test data. The top row indicates the Markov order of the H_0 that is tested. The numbers in the bottom two rows denote the power (fraction of correct rejections of the H_0) as well as the size (fraction of incorrect rejections of the H_0) of the χ^2 test.*

H_0	**0**	**1**	**2**	**3**	**4**	**5**
power	0.53	0.56	0.43	0.14	0.00	0.00
size	0.00	0.00	0.00	0.00	0.00	0.00

The most striking result visible in Table 1 is the very small size of the χ^2 test. Since we use a nominal significance level of 0.05 we would expect the size to be of a similar magnitude as well. The small size of the χ^2 test for Markov order has also been described in literature by Yakowitz [24]. The small size may not be a finite sample effect since the size remains small even for data of length 8192, which may also indicate a slow convergence to the χ^2 distribution of the test statistic. Another explanation for the low size may be that it is caused by correlations between the (overlapping) tuples extracted from the data used to estimate the transition probabilities. When we use less- or non-overlapping tuples to estimate the transition probabilities the size attains a value closer to the nominal value.

Another noticeable effect is the initial small increase of the power, followed by a more rapid decline of the power, for larger orders. The decrease of power for the highest orders was shown to be a finite sample effect since using larger sequences removed this effect. Also for sufficiently long sequences, we found that the power of the χ^2 test increases for increasing Markov order n.

5.4 Symbolic Surrogate Data Test

5.4.1 Typical Realization Surrogates

The power and size of the typical realization symbolic surrogate data test was investigated using the same test data as above. For each test sequences 19 typical realization surrogates were generated allowing for rejections of the H_0 at significance levels of at most 0.05. The results are displayed in Table 2. The top row indicates the Markov order of the H_0 data whereas the bottom rows denote power and size.

Table 2: *Estimation of the power and size of the symbolic typical realization surrogate test for Markov order. The top row indicates the order of the H_0 whereas the bottom rows represent the size and the power of the test.*

H_0	**0**	**1**	**2**	**3**	**4**	**5**
Power	0.49	0.53	0.55	0.47	0.30	0.09
Size	0.00	0.00	0.00	0.00	0.00	0.00

Clearly visible in Table 2 are on the one hand the very low size of the typical realization symbolic surrogate test and on the other hand the decrease of power for highest orders. The latter phenomenon is again a finite sample effect which disappears when using sufficiently long data sets. In that case we also see a persistent increase of the power for increasing Markov order of the H_0 similar to what was observed for the χ^2 test. The small size is reminiscent of effects described by Theiler and Prichard [22] who observed a small size for the test for nonlinearity using typical surrogate data generated by the "best" model fitted to the data. This was contributed to a lack of conditioning (constraining in their terminology). We will show that conditioning on the nuisance parameters in our problem also increases the power while yielding a true size close to the nominal size.

5.4.2 Constrained Probabilities

In this section the power and size of the constrained probabilities symbolic surrogate test was investigated in a very similar manner as done for the typical realization symbolic surrogates. For each of the 1000 sequences used in the previous sections we generated 19 constrained surrogate probabilities.

The results are displayed in Table 3. The top row indicates the Markov order of the H_0 whereas the bottom rows denote size and power of the constrained probabilities symbolic surrogate test.

Table 3: *Estimation of the power and size of the constrained probabilities surrogate test for Markov order. The top row indicates the order of the H_0 whereas the bottom* 6 *rows represent the size and the power of the test.*

H_0	**0**	**1**	**2**	**3**	**4**	**5**
Power	0.76	0.85	0.90	0.92	0.90	0.86
Size	0.06	0.05	0.04	0.06	0.05	0.05

From Table 3 we see that conditioning on the sample transition probabilities under the H_0 indeed results in an increase in size as well as values of the true size which are very close to the nominal size. The order at which the power starts to decrease due to finite sample effects has shifted somewhat to higher powers and is not very pronounced. Again, sufficiently long data sets completely remove this phenomenon.

In some cases it was found that the surrogate probabilities and realizations yielded conditional entropies which were identical to those of the original data. This is due to the limited number of possible combinations of sample probabilities used in the calculation of the conditional entropy, especially when only a relatively small number of of probabilities are used in those calculations (i.e. at low Markov orders). This effect was most pronounced when determining the size of the test for 0^{th} order Markov order (approximately. 5% of the surrogates) and disappeared when using sufficiently long sequences. When the value of the conditional entropy of the original data and the surrogates were equal we did not increase the rank of S^d.

5.4.3 Constrained Realizations

In this section the power and size of the symbolic constrained realization test was investigated. For each of the 1000 sequences used in the previous sections we generated 19 constrained surrogate realizations. The power of this test was estimated testing the H_0 of n^{th} Markov order by constructing n^{th} order surrogates, using test sequences of Markov order $n+1$ and h_{n+1} as test statistic. The size of the test is estimated by testing the H_0 of n^{th} Markov order by constructing n^{th} order surrogates, using n^{th} order test sequences and h_{n+1} as test statistic.

The results are displayed in Table 4. The top row indicates the Markov order of the H_0 whereas the bottom rows denote size and power of the symbolic typical realization surrogate test.

Table 4: *Estimation of the power and size of the constrained realization surrogate test for Markov order. The top row indicates the order of the H_0 whereas the bottom 6 rows represent the size and the power of the test.*

H_0	**0**	**1**	**2**	**3**	**4**	**5**
Power	0.75	0.84	0.91	0.92	0.92	0.87
Size	0.04	0.05	0.05	0.06	0.06	0.05

We see that the results in Table 4 do not differ significantly to those for the constrained probabilities presented in Table 3 previously. The conditioning on the sufficient statistic - the initial conditions and the sample transition probabilities - results in an increased power compared to the typical surrogate method and the expected size. Here too, we find an ties in the values of the conditional entropy for the original and surrogate data. The number is approximately equal to the number found for the constrained probabilities and decreases rapidly for increasing order and for sufficiently long sequences.

6 Discussion

In this study we proposed a novel test for the Markov order of a symbolic sequence. This method is conceptual similar to the method of surrogate data used in the testing for nonlinearity used for continuously valued time series and we have adopted some of the terminology used in that context to emphasize and clarify ideas regarding the role of conditioning. We compared variations of the symbolic surrogate test to the well known χ^2 test for Markov order and a method focusing on the convergence of the conditional entropy h_n using realizations of simple randomly generated Markov processes of known order.

The use of conditional entropy for estimating the Markov order suffers from finite sample effects and is inadequate to make statistical inferences from. We estimated the power and size of the χ^2 test for Markov order as well as of the different symbolic surrogate methods. The results show that both the χ^2 and typical realization symbolic surrogate test have low power for large Markov order due to finite sample effects. Conditioning on nuisance parameters in this problem results in considerable improvement in the performance of the symbolic surrogate test. The constrained symbolic surrogate tests have larger power, a true size close to the nominal size and better finite sample properties compared to the χ^2 and typical realization symbolic surrogate test.

Thus, the constrained symbolic surrogate method is a useful new tool for statistical inference on symbolic data sets and may be particularly useful when analyzing experimental data sets which are often short and autocorrelated. Although many experimental data are not discreet and the choice of a partition determines too a large extend the properties of the resulting symbolic sequence, we argue that one can in some cases make a meaningful partition of the sample space so as to highlight certain features of the data set or system under consideration.

Acknowledgments

The constrained surrogate probabilities were suggested by F. Takens.

References

[1] T.W. Anderson and L.A. Goodman
Statistical inference about Markov chains. *Ann. Math. Statist.* **28**, pp. 89–110, 1957.

[2] P. Billingsley
Statistical Inference for Markov Processes University of Chicago Press, Chicago, 1961.

[3] P. Billingsley
Statistical methods in Markov chains. *Ann. Math. Stat.* **32** pp. 12–40, 1961.

[4] D.R. Brillinger
Nerve cell spike train data analysis: a progression of technique. *J. Amer. Statist. Assoc.* **87** pp. 270–271, 1992.

[5] C. Chatfield
Statistical inference regarding Markov chain models. *Appl. Stat.* **22** pp. 7–30, 1973.

[6] D.R. Cox and P.A.W. Lewis
Statistical Analysis of Series of Events. Monographs on Applied Probability and Statistics, Chapman and Hall, London, 1966.

[7] J.P. Crutchfield and N.H Packard
Symbolic dynamics of noisy chaos. *Physica D* **7** pp. 201-223, 1983.

[8] W. Ebeling and A. Neiman
Long-range correlations between letters and sentences in texts. *Phys. A* **215** pp. 233–241.

[9] J.-P. Eckmann and D. Ruelle
Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.* **57**(3), pp. 617–656, 1985.

[10] B. Efron and R.J. Tibshirani
An introduction to the bootstrap Monographs on Statistics and Applied Probability 57, 1993, Chapman and Hall, New York.

[11] H. Herzel, A.O. Schmitt and W. Ebeling
Finite sample effects in sequence analysis *Chaos, Solitons and Fractals* **4**(1) pp. 97–113, 1994.

[12] P.G. Hoel
A test for Markoff chains. *Biometrika* **41**, pp. 430–433, 1954.

[13] W. Li and K. Kaneko
Long-range correlation and partial $\frac{1}{f^{\alpha}}$ spectrum in a noncoding DNA sequence. *Europhys. Lett.* **17** pp. 655–660, 1992.

[14] R.W. Katz
On some criteria for estimating the order of a Markov chain. *Technometrics* **23**(3), pp. 243–249, 1981.

[15] A.W. Macrae
On calculating unbiased information measures. *Psych. Bull.* **75**(4), pp. 270–277, 1971.

[16] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simmons and H.E. Stanley
Long-range correlations in nucleotide sequences. *Nature* **356** pp. 168, 1992.

[17] W.H. Press, B.P. Flannery, S.A. Teukolsky and W.T. Vetterling
Numerical Recipes in Pascal, 1989, Cambridge University Press, Cambridge, U.S.A.

[18] N. Reid
The roles of conditioning in inference. *Statist. Sci.* **10** pp. 138–157, 1995.

[19] Claude E. Shannon and Warren Weaver
The mathematical Theory of Communication, University of Illinois Press, 1963.

[20] X.Z. Tang, E.R. Tracy, A.D. Boozer, A. deBraauw and R. Brown
Symbol sequence statistics in noisy chaotic signal reconstruction. *Phys. Rev. E* **51**(5) pp. 3871–3889, 1995.

[21] J. Theiler, S. Eubank, A. Longtin, B. Galdrikan and J.D. Farmer
Testing for nonlinearity in time series: the method of surrogate data. *Physica D* **58** pp. 77–94., 1992.

[22] J. Theiler and D. Prichard
Constrained-realization Monte-Carlo method for hypothesis testing. *Physica D* **94**, pp. 221–235, 1996.

[23] H. Tong
Determination of the order of a Markov chain by Akaike's information criterion. *J. Appl. Prob.* **12** pp. 488–497, 1975.

[24] S.J. Yakowitz
Small-sample hypothesis tests of Markov order, with application to simulated and hydrological chains. *J. Amer. Statist. Assoc.* **71** pp. 132–136, 1976.