Is Bach’s brain a Markov chain? Recurrence quantification to assess Markov order for short, symbolic, musical compositions

Jack Murdoch Moore, Débora Cristina Corrêa, and Michael Small

Citation: Chaos 28, 085715 (2018); doi: 10.1063/1.5024814
View online: https://doi.org/10.1063/1.5024814
View Table of Contents: http://aip.scitation.org/toc/cha/28/8
Published by the American Institute of Physics

Articles you may be interested in

Recurrence network modeling and analysis of spatial data

Multivariate weighted recurrence network analysis of EEG signals from ERP-based smart home system

Extended recurrence plot and quantification for noisy continuous dynamical systems
Chaos: An Interdisciplinary Journal of Nonlinear Science 28, 085722 (2018); 10.1063/1.5025485

Deriving inferential statistics from recurrence plots: A recurrence-based test of differences between sample distributions and its comparison to the two-sample Kolmogorov-Smirnov test

Recurrence quantification analysis for the identification of burst phase synchronisation

Phase space reconstruction for non-uniformly sampled noisy time series
Is Bach’s brain a Markov chain? Recurrence quantification to assess Markov order for short, symbolic, musical compositions

Jack Murdoch Moore,1 Débora Cristina Corrêa,1,2, a) and Michael Small1,2,3

1School of Mathematics and Statistics, The University of Western Australia, Crawley, Western Australia 6009, Australia
2Complex Data Modelling Group, Faculty of Engineering, Computing and Mathematics, The University of Western Australia, Crawley, Western Australia 6009, Australia
3Mineral Resources, CSIRO, Kensington, Western Australia 6151, Australia

(Received 5 February 2018; accepted 26 June 2018; published online 27 August 2018)

It is rarely possible to precisely characterise the system underlying a series of observations. Hypothesis testing, which involves assessing simple assumptions about driving mechanisms, provides hope that we can at least rule out certain possibilities regarding the nature of the system. Unfortunately, the brevity, nonstationarity, and symbolic nature of certain time series of interest undermines traditional hypothesis tests. Fortunately, recurrence quantification analysis (RQA) has an established record of success in understanding short and nonstationary time series. We evaluate the suitability of measures of RQA as test statistics in surrogate data tests of the hypothesis that ten compositions by the Baroque composer J. S. Bach (1685–1750) arose from a Markov chain. More specifically, we estimate the size (the rate at which true hypotheses are incorrectly rejected) and power (the rate at which false hypotheses are correctly rejected) from empirical rejection rates across 1000 realisations, for each of the ten compositions, of the surrogate algorithm. We compare hypothesis tests based on RQA measures to tests based on the conditional entropy, an established test statistic for surrogate data tests of Markov order, and find that the RQA measure $L_{\text{max}}$ provides more consistent rejection of the fairly implausible hypothesis that Bach’s brain was a Markov chain. Published by AIP Publishing. https://doi.org/10.1063/1.5024814

I. INTRODUCTION

Humans may have been appreciating music since before our species existed.1 In the modern era, even the melodically-challenged but scientifically-minded can enjoy this ancient art. Diverse disciplines, including neuroscience,2 psychology,3 network science,4 and signal processing,5 seek a quantitative understanding of music. The complexity of the traditional source of melody, the human mind, motivates application of tools from nonlinear time series analysis, a mathematical discipline which seeks to understand and quantify data from complex sources.

Before unleashing on unknown data the complete panoply of nonlinear time series techniques, it would be worthwhile considering the possibility that the data arose in some uncomplicated, and perhaps convenient, manner. A hypothesis test allows assessment of the possibility that a time series of interest arose through a given, relatively simple process. The attraction for scientists of exotic or startling conclusions motivates hypothesis tests for the source of observations. According to Theiler and Prichard,6 the aim of such hypothesis tests is “to find the least interesting explanation that cannot be invalidated by the data.”

A simple hypothesis for symbolic data is that they arose from a discrete Markov chain. A discrete Markov chain, which will be defined in Sec. II A, is a random process with a limited memory. The extent of this memory, the number of values present or immediately past which in any way affect the probability distribution of the next value in the sequence, is called the order of the Markov chain.

What sort of process underlies music composition? We know that the perception of music is mainly controlled by expectation, which in turn is influenced by the presence of recurrent structures. In this study, we investigate the nature of Bach’s compositions by interpreting them as empirical symbolic data arising from a Markov chain, a simple sort of random process which is characterised by its finite memory. Under this hypothesis, a random walk on the Markov chain will generate data with properties similar to those of the original music, which arose from the mind of Bach. Reiteration’s important role in music is exploited via recurrence quantification analysis (RQA), a method of data analysis based on measuring patterns in the recurrence of states. We use RQA to characterise both the original composition and random walks on the Markov chain from which the original composition would arise with the greatest likelihood, permitting comparison of the two sorts of time series and thus allowing assessment of the possibility that a composition arose via a random walk. The melodies are short and non-stationary and thus are time series of a sort with which traditional techniques struggle but with which RQA has a history of success. Our study builds upon this precedent by demonstrating that, compared with an established technique based upon a measure of uncertainty called the conditional entropy, RQA provides more reliable rejection of the hypothesis that Bach’s remarkable brain was merely a higher order Markov chain.

a)Electronic mail: debora.corre@uwa.edu.au
Markov order is an important modelling parameter. In fact, correct estimation of the “memory” of a process is key to successful use of Markov chains in the modelling and forecasting of dynamic process belonging to diverse domains, including computational biology, language processing, and human page navigation modelling. Not surprisingly, Markov order has also been investigated during analysis of temporal dynamics of notes in a melody.

Due to its importance, inferring the order of a Markov chain from observed data has received much attention (see, for instance, Papapetrou and Kugiumtzis and their references). Test statistics for discriminating between rival Markov orders follows, in the limit of infinitely many observations, well-understood distributions. For example, a distribution asymptotically describes the negative log of the square of the proportional change in maximum likelihood as the assumed Markov order decreases from the true order. In order to avoid applying asymptotic assumptions to small data sets, Van der Heyden et al. test for Markov order using the method of surrogate data. The method of surrogate data involves generating artificial data which, under the null hypothesis, have properties similar to the original observations which we wish to understand. These artificial data are used to estimate the distribution under the null hypothesis of a chosen test statistic. For their tests of Markov order, Van der Heyden et al. used as test statistic the conditional entropy, which we define in Sec. II A.

Musical data incur challenges additional to those associated with other symbolic data. Musical sequences can be short and nonstationary, and additional experimentation cannot augment available data. We cannot ask J. S. Bach to pen for us an extra Sinfonia to decrease the statistical uncertainty in our analysis. Recurrence plots (RPs) have a history of success in illuminating data too short or nonstationary for more traditional methods. Recurrence plots have even been used to represent simultaneously the local and global recurrence structure of compositions. For instance, Serra et al. applied a variant of the RP, called a cross recurrence plot, to identify different versions (covers) of the same song, which suggests that RPs capture something essential about a composition. These successes make measures which quantify the patterns in RPs attractive discriminating statistics for hypothesis tests of the source of music data. The application of measures which quantify the patterns in RPs is called recurrence quantification analysis (RQA) and will be introduced in Sec. II C.

Dirst and Weigend consider the Fugue, a style of composition for which Bach is well known and which is closely related to the styles, the Invention and the Sinfonia, upon which our study focusses. They described the structure of Fugues in terms of the arrangement of musical themes (patterns of notes) and transformed versions of the same themes. Because both the amount of repetition and the amount of novel content in a composition play an important role in music experience, these themes and their images under transformations are combined with non-thematic music material to create satisfying harmony and rhythm. Although Dirst and Weigend did not believe that Fugues emerge from random combinations of the theme and its variations, their explanation of music in terms of “tension between randomness and order” highlights the dual importance of stochasticity and recursion to understanding the process of composition.

Since each set of surrogate data is a realisation of a Markov chain, it is natural to think of them as random walks on weighted networks. McCullough et al. used RQA measures in investigating the extent to which random walks on a network preserve the character of the time series from which the network arose. Specifically, McCullough et al. considered random walks on weighted ordinal transition networks constructed from time series derived from the Lorenz and other dynamical systems. Campanharo et al. and Hou et al. performed analogous investigations of random walks on non-ordinal transition networks, but without employing RQA.

Because the human experience of music depends upon expectation and recurrence, RQA measures have intrinsic significance in descriptions of music. More generally, the relative success of different hypotheses of memory may reflect musical composition conventions.

In this contribution, we benchmark using artificial data the reliability of the method of surrogate data using as test statistics four different RQA measures in addition to the conditional entropy. The results allow assessment of the reliability with which a person analysing a composition by J. S. Bach would be able to conclude that the source of the time series (the mind of Bach) was not a Markov chain.

Section II briefly introduces concepts, related to Markov chains, the method of surrogate data and recurrence plots, upon which subsequent analysis relies. Section III outlines the approaches used to benchmark hypothesis tests based upon different statistics and to investigate the Markov order of compositions by J. S. Bach. Section IV presents and interprets the implementation of these approaches. Finally, Sec. V summarises the main findings and suggests directions for future work.

II. BACKGROUND

In this section, we introduce briefly important concepts associated with Markov chains, hypothesis testing, and recurrence plots.

A. Markov chains

A random sequence indexed by time , each element of which takes values in a countable set , is called a Markov chain when there exists such that the probability distribution of values at each time depends only upon the value of the variable at the times which immediately precede time . For fixed , the necessary condition for a time-homogeneous Markov chain is that

where

respectively, denote the transition probabilities
The order of a Markov chain is the smallest \( n \) for which this property holds.

A Markov chain of order \( n \) is specified completely by its transition probabilities from words of length \( n \),
\[
\{ p( i_t | i_{t-1}, \ldots, i_{t-n} ) \ | \ A_{i_t}, A_{i_{t-1}}, \ldots, A_{i_{t-n}} \in S \}.
\]
That is, these transition probabilities parameterise the space of Markov chains of order \( n \) which take values in the set \( S \) of symbols.

The quantity which characterises the uncertainty in the value of an element of a Markov chain when the preceding \( m \) values are known is the conditional entropy of order \( m \),
\[
h_m \equiv - \sum_{(i_1 i_2 \ldots i_{m+1}) \in \mathbb{Z}^m} p( i_{m+1}, i_m, \ldots, i_1 ) \times \log p( i_{m+1} | i_m, \ldots, i_1 ),
\]
where \( p(i_{m+1}, i_m, \ldots, i_1) \) denotes the probability \( P( a_{m+1} = A_{i_{m+1}}, a_m = A_{i_m}, \ldots, a_1 = A_{i_1} ) \). For a Markov chain of order \( n \), it follows from (1) that
\[
\text{if } m \geq n \text{ then } h_m = h_n. \tag{2}
\]
This result is important for hypothesis tests of Markov order which employ as a test statistic the conditional entropy. In particular, in Sec. III C, we will employ the preceding inequality. In that section, we will also rely upon the following inequality, which applies to all random sequences rather than to Markov chains alone, and which Van der Heyden \textit{et al.} attributed to Shannon and Weaver:
\[
\text{for } m \geq n, \quad h_m \leq h_n. \tag{3}
\]

\section*{B. Hypothesis testing and the method of surrogates}

Analysis predicated upon a complicated assumption should be preceded by verification that a simpler assumption is not justified. Hypothesis testing offers a means of assessing simple assumptions.

Testing a null hypothesis \( H_0 \) can involve determining the probability, assuming \( H_0 \), with which a statistic \( X \) would be even more extreme than the value \( x \) of the statistic which was actually observed. This probability, called the \( p \)-value, can be calculated from the cumulative distribution function (CDF) of the statistic \( X \) under \( H_0 \). If the computed \( p \)-value is less than or equal to a previously determined probability \( \alpha \), called the size of the test, then \( H_0 \) can be rejected at the level of significance \( \alpha \). Traditional values of \( \alpha \) are \( \alpha = 0.05 \) and \( \alpha = 0.01 \).

The size \( \alpha \) of the test is the probability of rejecting \( H_0 \) assuming that this hypothesis is actually correct. This type of mistake is called type I error\textsuperscript{32} and so size is also known as the rate of type I error. The other sort of mistake which a hypothesis test can incur is type II error\textsuperscript{33} failure to reject \( H_0 \) even though \( H_0 \) is false. The rate of type II error is often written as \( \beta \), and it is also known as sensitivity. The power of a test is equal to \( 1 - \beta \) and is the probability of rejecting \( H_0 \) when \( H_0 \) is indeed false. The closer to unity is the power, the better would the test be able to discriminate against a false null hypothesis. Of course, \( \beta \), and hence the power of a test, depends upon the true nature of the observation.

The method of surrogates is a form of hypothesis testing which is particularly useful when the distribution of the statistic \( X \) under \( H_0 \) cannot be determined analytically. The method of surrogates uses some number \( B \) of instances of the statistic \( X \) randomly realised under \( H_0 \), together with the value \( x \) originally observed to estimate the CDF of \( X \). This estimate is
\[
F_X : \mathbb{R} \to [0, 1] \text{ given by}
\]
\[
F_X(y) = \frac{1}{B+1} \sum_{i=0}^{B} \Theta(x_i - y), \tag{4}
\]
where for \( i = 1, 2, \ldots, B, x_i \) denotes realisation \( i \) of \( X, x_0 = x \), and \( \Theta \) is the Heaviside step function \( \Theta : \mathbb{R} \to [0, 1] \) defined by \( \Theta(y) = \begin{cases} 0 & \text{if } y < 0 \\ 1 & \text{if } y \geq 0 \end{cases} \).

If only smaller values of a real-valued test statistic \( X \) are considered more extreme, then the test is called one-sided and the \( p \)-value is \( p = r/(B+1) \). In this expression, \( r \) is the rank of \( x \), which is the position of \( x \) when \( x \) and the other \( B \) realisations of \( X \) are ranked from least to greatest. It follows that the null hypothesis \( H_0 \) can be rejected when the original statistic \( x \) lies among the uniquely smallest \( \lfloor \alpha(B+1) \rfloor \). If both the smallest and largest values of a real-valued test statistic \( X \) are considered extreme, then the test is called two-sided and the \( p \)-value is \( p = 2 \min[r/(B+1), (B+2-r)/(B+1)] \). It follows that the null hypothesis \( H_0 \) can be rejected when the original statistic \( x \) lies among either the uniquely largest or the uniquely smallest \( \lceil \frac{2}{3}(B+1) \rceil \).

The requirement for rejection of the null hypothesis that the value of the statistic observed in the original data be strictly more extreme, rather than as or more extreme than all but \( \lfloor \alpha(B+1) \rfloor - 1 \) surrogates, is slightly conservative. If there were ties then, even were the empirical distribution of the test statistic identical with that of the null hypothesis, the true size \( \alpha \) could be less than the nominal size \( \alpha \). Van der Heyden \textit{et al.} avoided this potential bias by assigning permissible ranks uniformly at random among realisations whose values coincided. Our approach has the advantage that it allows the method of surrogates to be interpreted as proceeding from an empirical estimation via (4) of the cumulative distribution function of the statistic \( X \).

\section*{C. Recurrence plots and recurrence quantification analysis}

Consider a series of states \( s(1), s(2), \ldots, s(N) \) and a threshold \( \epsilon > 0 \) which corresponds to the distance by which two states must be separated to be distinguishable from one another. When distance \( d[s(i), s(j)] < \epsilon \), we say that the state at time \( i \) recurred at time \( j \) or say that the states at \( i \) and \( j \) are similar. The \( N \times N \) matrix \( R = (R_{ij}) \) defined by
\[
R_{ij} = \begin{cases} 1 & \text{if } d[s(i), s(j)] \leq \epsilon \\ 0 & \text{if } d[s(i), s(j)] > \epsilon \end{cases}
\]
is called the similarity or recurrence matrix and captures the patterns of recurrence. These patterns can be visualised using a recurrence plot (RP), which is an \( N \times N \) grid constructed from a recurrence matrix by filling a square at coordinates \( (x,y) = (i,j) \) if and only if \( R_{ij} = 1 \). The line \( i = j \) is called
the line of identity (LOI). When the distance is a metric, for which always \(d(s, s) = 0\), the LOI is always filled.

If diagonally adjacent coordinates \((i, j)\) and \((i+1, j+1)\) both correspond to recurrent states, then it means that the similar states \(s(i)\) and \(s(j)\) evolved to states \(s(i+1)\) and \(s(j+1)\) which were also similar. Thus, diagonal lines in a recurrent plot are a hallmark of deterministic evolution according to a rule which depends continuously on the state. In a periodic system, similar states at times \(i\) and \(j\) can remain similar under forward and reverse time evolution, filling a diagonal line which passes through the coordinate \((i, j)\) and spans the entire RP. In contrast, in deterministic chaos, in which nearby states diverge, diagonal lines on the recurrence plot have limited length. Thus, the distribution in a RP of diagonal lines is rich with meaning. Other features of RPs, such as vertical lines, horizontal lines, and composite squares, have other interpretations which are outlined, for example, by Marwan et al.\(^{34}\)

Eckmann et al.\(^{35}\) introduced RPs as a tool for visual characterisation of dynamical systems. Zbilut and Webber\(^{36}\) provided an alternative to qualitative evaluation of the features of RPs when they introduced quantitative measures of RP structure. The utilisation and interpretation of these and similar measures is known as recurrence quantification analysis (RQA). For this contribution, we consider four RQA measures, which we calculated using the CRP Toolbox for MATLAB\(^{37}\). Marwan et al.\(^{34}\) provide more information about the toolbox, as well as a more extensive list of RQA measures.

Because of the importance in music of reiteration,\(^{31}\) we focus upon four RQA measures which quantify the occurrence in a RP of diagonal lines. Hence these RQA measures can be defined in terms of the probability distribution \(P : \mathbb{N} \to [0, 1]\), \(l \mapsto P(l)\) of lengths \(l\) of diagonal lines.

The maximum diagonal line length \(L_{\text{max}}\) is the longest observed maintenance of recurrence. Visually, the maximum diagonal line length is the longest diagonal line other than the LOI.

The determinism

\[
\text{DET} = \frac{\sum_{l=\min}^{\infty} lP(l)}{\sum_{l=1}^{\infty} lP(l)} \tag{5}
\]

is the probability that a randomly chosen pair of similar states will be part of a recurrence of length at least \(l_{\min}\). Visually, the determinism is the fraction of recurrent points which lie on lines of length at least \(l_{\min}\).

The mean length of diagonal lines

\[
\langle L \rangle = \frac{\sum_{l=\min}^{\infty} lP(l)}{\sum_{l=\min}^{\infty} P(l)}
\]

is the mean length of time for which a new recurrence will be maintained, assuming that it lasts for at least \(l_{\min}\) time steps.

In the context of recurrence plots, the entropy usually refers to the Shannon entropy of the distribution of line lengths

\[
\text{ENTR} = -\sum_{l=\min}^{\infty} P(l) \log[P(l)].
\]

Intuitively, the entropy is the uncertainty in the length of time for which a new recurrence of length at least \(l_{\min}\) will be maintained. If recurrences of length at least \(l_{\min}\) are distinguished only by the length of time for which they last, then \(\text{ENTR}\) is the uncertainty in type of a recurrence which has already been maintained for time \(l_{\min}\).

A recurrence plot can also be visualised as a type of complex network called a recurrence network.\(^{37}\) The nodes of a recurrence network comprise distinct values of the time series, and distinct nodes \(s(i)\) and \(s(j)\) are linked if and only if \(R_{ij} = 1\). That is, nodes \(s(i)\) and \(s(j)\) are linked if and only if the state at time \(i\) recurred at time \(j\). This interpretation of a RP makes available to RQA the metrics of complex networks.\(^{37,38}\)

Although in Sec. IV, we understand and visualise results using a complex network representation of a musical time series, the directed links of the transition networks\(^{37}\) which feature in Sec. IV are defined quite differently from the undirected links of a recurrence network. Our current study does not make use of recurrence network metrics, but they might prove useful test statistics in future work.

III. METHODS

Our nominal purpose was to see whether or not Bach’s compositions could be distinguished from realisations of Markov chains. More broadly, we were valuating the effectiveness of different discriminating statistics in assessing the nature of short, symbolic time series containing a number of symbols which were relatively large in comparison to the number of data. We considered using in the method of surrogate data five different discriminating statistics. These five are the conditional entropy of Sec. II C and the four RQA measures defined in Sec. II A; \(L_{\max}\), \(\text{DET}\), \langle \text{L} \rangle, \text{ and } \text{ENTR}\). Van der Heyden et al.\(^{19}\) used the method of surrogate data to test hypotheses that short, symbolic time series had Markov order \(n\) in the range \(0 \leq n \leq 5\). They employed as their test statistic the conditional entropy of order \(n\).

Recurrence plots do depend upon the metric with respect to which they are defined, but it was not clear what metric would assign the most meaningful distance between each pair of musical notes. Since the data were discrete, we used the discrete metric defined by \(d(A, B) = \begin{cases} 0 & \text{if } A = B \\ 1 & \text{if } A \neq B \end{cases}\)

some fixed threshold \(\epsilon \in [0, 1]\). A Markov chain of order \(n\) produces correlations of length \(n\). To focus on correlations not explicitly enforced by the null hypothesis, tests for the hypothesis of Markov order \(n\) used \(l_{\min} = n + 1\). Recurrence quantification analysis with \(l_{\min} = 2\) on data embedded in dimension \(n + 1\) produced similar results, as would, likely, many alternative choices of RQA parameters.

A. Selection and preparation of data

We considered a total of ten keyboard compositions by the composer Johann Sebastian Bach (1685–1750);\(^{39}\) five Inventions and five Sinfonias. Bach’s Inventions (also known as two-part Inventions) are compositions with two leading voices, while his Sinfonias (also known as three-part Inventions) have three leading voices. Although the different voices...
are designed to be played simultaneously, each is an independent melody and contains by itself the main characteristics of the composition.\textsuperscript{24,25} We derived our time series from the first leading voice of compositions. We chose compositions which would lead to symbolic time series of similar length; each Invention led to a time series containing between 228 and 255 symbols, while each Sinfonia led to a time series containing between 221 and 259 symbols. The Appendix provides additional details.

These ten compositions were downloaded in MIDI format from IMSLP.\textsuperscript{40} We used the MIDI Toolbox\textsuperscript{41} to transform these recordings into symbolic time series of discrete notes. In our analysis, each note of a composition comprises a pitch-duration pair. The pitch is the frequency of the sound, while the duration is the length of time between successive notes. A melodic pattern can be repeated at a different tonality, meaning that the pattern recurs translated by some number of units of pitch. In order to recognise recurrence of melodic patterns even when they occurred across different tonalities, throughout each composition, we transformed the notes such that they were defined with respect to a constant tonality, either the C major or the A minor tonality. Similarly, we set the beat unit equal to a quarter note to allow recognition of similar rhythmic patterns occurring at different duration ratios

Each pitch-duration pair (each note), distinct modulo tonality, corresponded to a distinct symbol (a distinct node in the network). Our analysis included none of the many other characteristics of music, such as, for example, intensity or timbre.\textsuperscript{42}

### B. Surrogate algorithm

The method of surrogate data involves generating new data sets which under the hypothesis $H_0$ have properties similar to the data of interest. To this end, when evaluating the null hypothesis that the data $(a_i)_{i=1}^N$ arose from a Markov chain of order $n$, we generated $B$ sets of $N$ surrogate data $(a_i^{(B)})_{i=1}^N$ in the following way. To begin with, we appended to the $N$ original data its first $n$ entries, forming the $N+n$ appended data $(a_i)_{i=1}^{N+n}$. The first $n$ entries of the sequence were appended to the end of the original time series such that $a_{N+i} = a_i$ to ensure that no transition probability was left undefined for a word of length $n$ which appeared in the original $N$ data. Next, for each $i = 1, 2, \ldots, B$, we set $a_1^{(B)} = a_1, a_2^{(B)} = a_2, \ldots, a_n^{(B)} = a_n$ and randomly transitioned to the subsequent $N-n$ symbols $a_{n+1}^{(B)}, a_{n+2}^{(B)}, \ldots, a_N^{(B)}$ according to the transition probabilities $\{p(\hat{i}_1, i_{1-n}, \ldots i_{n-n})\}$ with which the appended original data $(a_i)_{i=1}^{N+n}$ would have arisen with maximum likelihood.

The $B$ new data sets arose from a Markov chain of order $n$ and so were suitable for testing the hypothesis that the original sequence $(a_i)_{i=1}^N$ arose from a Markov chain of this order. However, these surrogate data did not, in general, exhibit transition probabilities from words of length $n$ equal to those of other surrogate data generated through the same process, or equal to those of the original data. We highlight this because in the limit as the number $N$ of data grows large, we would expect the observed transition probabilities to approach those of the homogeneous and irreducible\textsuperscript{43} Markov chain from which the data were generated.

The maximum likelihood transition probabilities were straightforward to calculate. As Van der Heyden \textit{et al.}\textsuperscript{19} explained, for $n \geq 1$, the maximum likelihood transition probabilities are proportional to the observed transitions from each word of length $n$ which appears. When $n = 0$, the procedure simply sampled uniformly, with replacement, from the observed density of symbols.

### C. Calculation of rejection rates including size and power

We rejected the null hypothesis only when the original data gave rise to the value of the statistic which was uniquely most extreme with respect to all statistics which arose from the surrogate and the original data. Neglecting ties, the nominal size $\alpha_N$ of such a test is that calculated from the empirical CDF of the statistic. That is, $\alpha_N = \frac{1}{B+1}$ for a two-sided test and $\alpha_N = \frac{1}{B}$ for a one-sided test. We applied two-sided tests to RQA measures and, following Van der Heyden \textit{et al.}\textsuperscript{19} in taking advantage of the relationships (2) and (3) of Sec. II A, applied one-sided tests to conditional entropies. To maintain the nominal size $\alpha_N$ of the tests, if $B$ sets of surrogate data were used for RQA measures, then only the first $(B-1)/2$ sets generated were used for the conditional entropy.

There may be differences between the empirical and true CDFs of the test statistic $X$ under a null hypothesis, which can lead to differences between the nominal and true size of the test. We estimated the true size $\alpha$ of the test as the fraction of rejections of the null hypothesis that the process was a Markov process of order $n$ when surrogates were indeed generated from a Markov chain of order $n$. We estimated the power of the test as the fraction of rejections of the null hypothesis that the process is a Markov process of order $n$ when, in fact, surrogates were generated from a Markov chain of order $n+1$.

The ease of tuning the nominal size of a hypothesis test based on surrogate data makes desirable agreement between nominal and true size. Theiler and Prichard\textsuperscript{44} called "accurate" hypothesis tests for which the nominal and true size agree.

Rejection rates estimating the size and power were based upon application of the surrogate algorithm to 1000 realisations for each composition. Each realisation was a random realisation of a maximum likelihood Markov chain of the original music data, generated as though by the surrogate algorithm of Sec. III B. In particular, these realisations were of the same length as the original composition.

These estimates of size and power guide interpretation of the rejection rates which were the crux of this project: those calculated from application of the surrogate algorithm to each of the ten compositions by J. S. Bach. At each fixed nominal significance $\alpha_N$, the hypothesis of each Markov order was assessed only once per composition. Hence, these rejection rates were effectively a simple fraction with a denominator of ten. Those confident that Bach’s brain was not a Markov chain may interpret these rejection rates as another estimate of the power of the hypothesis test.
IV. RESULTS AND DISCUSSION

Figure 1 shows rates of rejection, using a test with nominal power $\alpha_N = 0.05$, of the null hypothesis that a time series arose from a Markov process of orders between 0 and 14. In particular, Fig. 1 suggests that RQA measures were more suitable than the conditional entropy for testing the hypothesis of Markov order 0. Figures 1(b) and 1(c) show that the conditional entropy $h_n$, along with all RQA statistics other than $DET$, led to similar rates of correct rejection of the null hypothesis. However, Fig. 1(a) shows that the size $\alpha$ of tests of the hypothesis of a Markov chain of order 0 based upon the conditional entropy $h_0$ far exceeded the nominal rejection rate $\alpha_N = 0.05$. In contrast, tests based upon RQA measures met the criterion $\alpha \leq \alpha_N$. Note that when $n = 0$, because $L_{\min} = n + 1 = 1$, the determinism $DET$, defined by (5), was always unity and so could not be expected to lead to rates of rejection other than zero.

For hypotheses of higher Markov order, the advantages of RQA measures were present, but were less prominent—Fig. 1(a) shows that for order greater than or equal to one, the conditional entropy often led to a true size significantly lower than that corresponding to any RQA measure. Figure 1(b) shows that surrogate data methods based upon RQA measures other than $DET$ exhibited, in benchmarking, power significantly less than that based on conditional entropy. In contrast, for orders one and two, $DET$ led to power similar to that corresponding to the conditional entropy and for orders between three and eight, $DET$ led to a higher power than did the conditional entropy. Figure 1(c) details the rates of rejection of hypotheses that original compositions by J. S. Bach arose from Markov processes. This figure shows differences between test statistics very different from those which arose in benchmarking. In practice, at a nominal rate $\alpha_N = 0.05$, RQA measures allowed, for considerably higher $n$, rejection of the null hypothesis that compositions arose from a Markov chain of order $n$. The conditional entropy allowed rejection of the hypotheses of the compositions having Markov order of one or less at a rate of 100%, but of Markov orders two or higher at rates of 10% or less. In contrast, three RQA measures—$(L), ENTR$, and, except at order zero, $DET$—allowed rejection of hypotheses of Markov order four or less at a rate of at least 50%.

Although for fixed nominal size $\alpha_N = 0.05$, Fig. 1(c) shows that the conditional entropy allows for original compositions less consistent rejection than do RQA measures of the hypothesis of Markov orders 2–6, Fig. 1(a) suggests that it does so with significantly less risk of false rejection; with a lower true size $\alpha$. Comparison of Figs. 2 and 3 reveals the advantage of RQA measures, and especially the advantage of $L_{\max}$. Figure 2 shows that a nominal size $\alpha_N = 0.01$ provided to hypothesis tests based upon $L_{\max}$ an estimated true size $\alpha$ less than 0.05 alongside a rate of rejection of at least 40% of null hypotheses of Markov order 3–5. Figure 3 shows that, in comparison, a nominal size $\alpha_N = 0.25$ led, for null hypotheses of some Markov orders in the range 3–5, to tests based upon the conditional entropy with an estimated true size $\alpha$ greater than 0.05 alongside a rate of rejection no more than 20%.

This advantage of $L_{\max}$ in distinguishing musical compositions from random time series is not particularly surprising. The repetition of patterns, with or without variation, is integral to many forms of music and is especially evident in Bach’s Inventions and Sinfonias. This property places RQA in general, and the measure $L_{\max}$ in particular, in a good position to distinguish the work of J. S. Bach from the output of a Markov chain.

As previously noted, Fig. 1(c) exhibits, when the conditional entropy is employed as the test statistic, a sharp
transition from a rate of rejection of 100% for hypotheses of Markov order one or less to a rate of rejection of 10% or lower for hypotheses of Markov order of two or more. Figure 1(b) shows for the same test statistic a far more gradual transition from a perfect rate of rejection for hypotheses of Markov order one or less to a rate of rejection of 10% or lower for hypotheses of Markov order of six. The failure to reject, for original musical time series, hypotheses of Markov order two or more may suggest that Markov order two sufficed to capture an important part of musical compositions; perhaps, an aspect associated with simple music composition conventions. It would be interesting to see whether the pattern persists for shorter or longer musical time series, for pieces by other composers, and for different styles.

The failure to reject, using any test statistic, hypotheses of order nine or greater is probably mostly due to the brevity of the time series and the large number of distinct symbols which they feature. (See Table I for quantification of these and other properties of the compositions.) This combination will tend to reduce the number of words of length \( n \) which appear more than once in the time series, thus reducing the number of words which can possibly be observed to precede more than one distinct word of length \( n \). This reduces the number of states of the Markov chain at which there is any randomness, and so could make it harder to produce surrogates which differ significantly from the original data.

Figure 4 illustrates the reduction in randomness as Markov order \( n \) increased. The figures depict weighted, directed networks which represent the Markov chains constructed from Invention 1. The nodes of the networks representing a Markov chain of order \( n \) are the words of length \( n \) in the time series from which the Markov chain is constructed. Nodes with outdegree \( d^+ \) greater than or equal to two represent a state from which more than one transition is possible. Figure 4(a) represents the Markov chain of order

| Table I. Details of the compositions, and the time series to which they correspond, which we have considered. In this table, \( N \) denotes the length of, and \( \nu_1 \) denotes the number of distinct values which appear in, the time series. |
|-----------------|----------|----------|----------|-------|-------|
| Style          | Number  | Catalogue | Key      | \( N \) | \( \nu_1 \) |
| Invention 1    | 1        | BWV 772   | C major  | 242   | 30    |
| Invention 2    | 2        | BWV 773   | C minor  | 249   | 31    |
| Invention 3    | 3        | BWV 774   | D major  | 246   | 28    |
| Invention 4    | 4        | BWV 775   | D minor  | 255   | 43    |
| Invention 5    | 5        | BWV 776   | Eb major | 228   | 20    |
| Sinfonia 1     | 1        | BWV 787   | C major  | 247   | 58    |
| Sinfonia 2     | 2        | BWV 788   | C minor  | 259   | 38    |
| Sinfonia 3     | 3        | BWV 789   | D major  | 238   | 55    |
| Sinfonia 4     | 4        | BWV 790   | D minor  | 226   | 46    |
| Sinfonia 5     | 5        | BWV 791   | Eb major | 221   | 45    |
FIG. 4. Networks constructed from Invention 1 and depicted using the network visualisation software Gephi. The nodes are the words of length: (a) $n = 1$ and (b) $n = 6$. The links are weighted according to the empirical probability of transitions from states, and the thickness with which they are drawn increases with weight. The links are directed and only traversed clockwise. Each surrogate for the hypothesis of Markov order $n$ comprised a random walk starting from the word of length $n$ which began the original composition.

$n = 1$ and contains $v_1 = 30$ nodes, 19 of which have outdegree $d^+$ greater than or equal to two. In contrast, Fig. 4(b), which represents the Markov chain of order $n = 6$, contains $v_6 = 229$ accessible states, but only 6 have outdegree $d^+$ greater than or equal to 2. Except at these 6 states, the surrogate algorithm was constrained to follow deterministic chains without any element of randomness. As order $n$ exceeded the maximum diagonal line length $L_{\text{max}}$, no word of length $n$ occurred more than once, and all randomness disappeared. The network which represented the Markov chain of order $n$ became a loop, and the “random” walk became a deterministic cycle. For Invention 1, $L_{\text{max}} = 7$ and determinism takes over from order $n = 8$.

States from which more than one transition was possible are also rare in Markov chains of order $n = 6$ constructed from any of the other four inventions. Figure 5 shows the variation with $n$ of the fraction of nodes with outdegree $d^+$ greater than or equal to two in networks representing Markov chains of order $n$ constructed from inventions by J. S. Bach. For the four inventions other than that from which Fig. 4 were constructed, for $n = 8$, the fraction of nodes with outdegree $d^+$ greater than one was small but, although it is difficult to discern from the figure, nonzero. This implies, in particular, that the other inventions have maximum diagonal line length $L_{\text{max}}$ greater than seven.

V. CONCLUSION

We have investigated the null hypothesis that Bach’s brain was actually a Markov chain. The RQA measure $L_{\text{max}}$ allowed rejection of null hypotheses that compositions by J. S. Bach arose from Markov chains more consistently than did the conditional entropy, an established test statistic for Markov order. The advantages of RQA measures were particularly evident when power was treated as a function of the nominal size $\alpha_N$. However, the RQA measure $L_{\text{max}}$ was also advantageous when power was assessed as a function of an estimate of the true power $\alpha$. The power of RQA measures in testing hypotheses of the origins of music data may arise from the importance to music of recurrence.

The failure to reject the hypothesis of a Markov chain of Markov order greater than or equal to two when using conditional entropy as a test statistic suggests that Markov chains of order two may capture an important aspect of composition conventions. Analogous patterns for Markov order nine or greater when using the RQA measure $L_{\text{max}}$ spring from the
shortness of, and the large number of distinct symbols featured in the time series constructed from the ten compositions which we have considered.

Van der Heyden et al. also used the method of surrogates to test Markov order. Their work, which used the conditional entropy as a test statistic, inspired and guided our contribution. Although our comparison of conditional entropy and different RQA measures is, in itself, interesting and useful, a valuable next step would be to follow Van der Heyden et al. further in using constrained, rather than unconstrained, surrogate methods. Constrained surrogate methods involve generating sets of surrogate data conditioned on their having, under the null hypothesis, a sufficient statistic coinciding with that of the original data. This practice increases the computational burden of hypothesis testing, but also improves accuracy and increases power.19

ACKNOWLEDGMENTS

M.S. is supported by the Australian Research Council through Discovery Project No. DP140100203. J.M.M. was supported by the UWA Prescott, the CSIRO OCE, and the CSIRO-UWA Complex Engineering Systems PhD scholarships.

APPENDIX: TEN COMPOSITIONS BY J. S. BACH

This appendix provides, via Table I, details of the ten compositions which we considered. For $j \geq 1$, let $v_j$ denote the number of words of length $n$ in the time series once it has been appended as described in Sec. III B, and let $v_0 = v_1$. With these choices, for each $j \geq 0$, the number of accessible states in the Markov chain of order $j$ built from the time series is equal to $v_j$. Figure 6 shows that as the order $n$ increases, the number $v_n$ of accessible states in the Markov chain of order $n$ approaches from below the length $N$ of the time series from which it was constructed. When the order $n$ exceeds the maximum diagonal line length $L_{max}$, $v_n = N$, and the Markov chain of order $n$ comprises a deterministic loop. Figure 6(a) illustrates this for Inventions, while Fig. 6(b) illustrates this for Sinfonias.

23 L. Dreyfus, Bach and the Patterns of Invention (Harvard University Press, 1996).