Explaining Interval Sequences by Randomization

Andreas Henelius, Jussi Korpela, and Kai Puolamäki

Finnish Institute of Occupational Health,
Topeliuksenkatu 41 a A, FI-00250 Helsinki, Finland
{andreas.henelius,jussi.korpela,kai.puolamaki}@ttl.fi

Abstract. Sequences of events are an ubiquitous form of data. In this paper, we show that it is feasible to present an event sequence as an interval sequence. We show how sequences can be efficiently randomized, how to choose a correct null model and how to use randomizations to derive confidence intervals. Using these techniques, we gain knowledge of the temporal structure of the sequence. Time and Fourier space representations, autocorrelations and arbitrary features can be used as constraints in investigating the data. The methods presented are applied to two real-life datasets; a medical heart interbeat interval dataset and a word dataset from a book. We find that the interval sequence representation and randomization methods provide a powerful way to explore interval sequences and explain their structure.

1 Introduction

Time series are sequences of consecutive, time-stamped events. The events can have properties, such as values of measurements at the particular time instances. A single event can have multiple properties, in which case one ends up with a multidimensional time series. In this paper we, however, ignore the properties of the events and study only the fundamental temporal structure of the time series, which can be represented as a sequence of intervals.

Interval sequences are ubiquitous. They can be analyzed and compared by numerous methods, and many application areas such as medical signal processing have established conventions on how to study them. The structure of event sequences can be described by complex models. However, before addressing more complex properties of the event sequence, the first question is whether it is meaningful to look for complex structures. Can the structure observed in the interval sequence be explained by a random occurrence? If not, then what constitutes a good description?

Randomization methods provide a means of studying non-random structures. These techniques have a long tradition in statistics and are used increasingly in data analysis as well. To use randomization methods one first needs to define the null distribution from which random samples are drawn. If the observed event sequence differs, in terms of one or more test statistics, from the random samples, we can conclude that there are non-random structures.

The null distribution encodes our prior information and assumptions about the data as constraints. The choice of constraints is, however, far from trivial and

H. Blockeel et al. (Eds.): ECML PKDD 2013, Part I, LNAI 8188, pp. 337–352, 2013.

no clear guidelines exist. A suitable set of constraints depends on the research question and hence there is no universally appropriate null model. A natural choice is to select constraints that explain those aspects of the data we assume known and wish to account for. This makes previously unknown patterns stand out. The randomization methodology can therefore be used as a probabilistically robust means of detecting statistically significant patterns [14,5,6,20].

1.1 Structure and Contributions of This Paper

In this paper we show how an event sequence can be represented as a sequence of events in the time and Fourier domains, and in the autocorrelation space. We present the main theoretic properties of these representations in Section 2. We introduce fast and convenient randomization methods in which various properties such as Fourier amplitudes or phases, autocorrelation coefficients, or arbitrary statistics of the event sequence are kept fixed. We demonstrate how these randomization techniques can be used to determine if the observed features of the event sequence are just random artifacts, and show how to detect the features explaining the event sequence. In Section 3 we demonstrate our approach by applying the methods to important real-life data consisting of heart rate variability data and the occurrence of words in natural language.

Summarizing, the main contributions of this paper are:

- Interval sequence representation and its theoretic properties.
- Efficient randomization techniques for interval sequences.
- Using randomization to derive confidence limits and to explain non-random features of the data.
- Application of the proposed methods in two real-life applications.

2 Methods and Theory

2.1 Definitions

Assume that we have a sequence of $N+1$ events that occur at times $t_0, t_1, \ldots, t_N$, where $t_0 \leq t_1 \leq \ldots \leq t_N$. In this paper, we consider a sequence of N intervals S, defined by

$$S = (x_0, x_1, \ldots, x_{N-1}),$$

where $x_n = t_{n+1} - t_n$. In most of the numerical formulæ we use the logarithmic interval sequence $S_z = (z_o, z_1, \ldots, z_{N-1})$, where $z_n = \log x_n$. The logarithmic scale is more appropriate for our two applications: doubling and halving the interval both cause equal absolute changes in the value of the logarithm of the interval sequence, whereas without use of the logarithm, long intervals would receive much larger weight in the analysis. Furthermore, logarithms of intervals can take any value, including negative, which is numerically convenient.

For convenience and where appropriate, we extend the interval sequence by assuming that it is cyclic with a cycle of length N, i.e., $x_{n+N} = x_n$ and $z_{n+N} =$

z_n for all n. We denote the mean by $\overline{z}$ and the variance by σ_z^2, defined by $\overline{z} = \sum_{n=0}^{N-1} z_n/N$ and $\sigma_z^2 = \sum_{n=0}^{N-1} (z_n - \overline{z})^2/N$, respectively.

Fourier Representation. The Fourier representation of the data is defined by the sine and cosine series,

$$\begin{aligned} z_n &= a_0 + \sum_{k=1}^{K} a_k \cos \frac{2\pi kn}{N} + \sum_{k=1}^{K} b_k \sin \frac{2\pi kn}{N} \\ &= a_0 + \sum_{k=1}^{K} c_k \cos\left(\frac{2\pi kn}{N} - \varphi_k\right), \end{aligned} \tag{1}$$

where $K = \lfloor N/2 \rfloor$. The data can be parametrized either by the parameters $(a_0, \{a_k\}, \{b_k\})$ or $(a_0, \{c_k\}, \{\varphi_k\})$, where $k \in \{1, \ldots, K\}$ The Fourier amplitudes satisfy $c_k = \sqrt{a_k^2 + b_k^2}$ and the Fourier phases satisfy $\varphi_k \in [0, 2\pi)$. The Fourier parameters and the inverse transformation can be computed in $O(N \log N)$ time by a Fast Fourier Transform (FFT).

Autocorrelation. We use the autocorrelation function r_l with lag l, defined by

$$r_l = \frac{1}{N} \sum_{n=0}^{N-1} \frac{(z_n - \overline{z})(z_{n+l} - \overline{z})}{\sigma_z^2}. \tag{2}$$

A value of the autocorrelation function for a single lag can be computed in $O(N)$ time, and the values of the autocorrelation function for all lags can be computed in $O(N \log N)$ time by using the fast Fourier transformation. Notice that due to the cyclicity assumption, the lags satisfy $r_l = r_{N-l}$; therefore, it is sufficient to consider lags in $l \in \{1, \ldots, \lceil N/2 \rceil\}$ only.

2.2 Randomization Methods

We define several distributions of interval sequences, and the respective randomization methods. Each of the distributions preserves some aspect of the original sequence.

Interval Randomization. The INTERVAL distribution is a uniform distribution over all permutations of sequence S. A sample S^* from the INTERVAL distribution can be drawn by permuting S uniformly in random.

Fixed Subsequence Randomization. The SUBSEQUENCE distribution is a uniform distribution over all permutations of the sequence S where a given subsequence $G_x \subseteq \{0, \ldots, N-1\}$ of the intervals is kept fixed. A sample S^* is obtained by permuting all intervals in S that are not in G_x uniformly in random.

Fixed Fourier Parameters Randomization. The FOURIER distribution is a distribution of interval sequences in which given subsets of Fourier amplitudes and phases have been fixed. The FOURIER distribution is obtained by fixing a subset $G_c \subseteq \{1, \ldots, K\}$ of Fourier amplitudes c_k where $k \in G_c$, and a subset $G_\varphi \subseteq \{1, \ldots, K\}$ of Fourier phases φ_k where $k \in G_\varphi$. A sample S^* from FOURIER is obtained by first taking a sample S'^* from the INTERVAL distribution, and then replacing the Fourier amplitudes c_k and Fourier phases φ_k not in G_c and G_φ, respectively, by the respective Fourier parameters of the sample S'^*. The sample S^* is then obtained by applying the inverse Fourier transformation of Equation (1) to the randomized Fourier parameters.

Uniform Randomization. As a comparison to the INTERVAL distribution, we define the UNIFORM distribution to a uniform distribution over all sequences of N intervals in which the duration is fixed to $t_N - t_0$.

Fixed Distance Function Randomization. Finally, we define a randomization method which approximately preserves any arbitrary constraint. We define the constraint by a *distance function* $d(S')$ which is a non-negative function of permutations of the original interval sequence and zero for the original non-permuted sequence, $d(S) = 0$. We define a distribution DISTANCE by

$$f(S') \propto e^{-d(S')}, \tag{3}$$

where S' is a permutation of the original interval sequence S. A sample from the distribution f is likely to include intervals which are close to the original interval sequence in terms of the distance function. A sample S^* from DISTANCE can be obtained via Markov chain Monte Carlo (MCMC) integration, described in more detail in Section 2.4.

We use the distribution DISTANCE to sample intervals preserving the autocorrelation function at lags given in $G_r \subseteq \{1, \ldots, \lceil N/2 \rceil\}$. We use the distance function $d(S') = \lambda \sum_{l \in G_r} |r'_l - r_l|$, where $\lambda > 0$ is a parameter describing the accuracy to which we want to preserve the autocorrelations and r'_l is the value of autocorrelations for the resampled sequence. There is a critical value of λ in Equation (3) corresponding to the phase transition in statistical physics: the threshold value is recognized from the fact that when λ exceeds the threshold most of the probability mass of f is close to the original interval sequence (i.e., the expected value of the distance function is small). For all datasets considered in this paper a sufficiently high value is $\lambda = 10^4$, which is used in all experiments. Notice that the MCMC method could also be used to preserve the Fourier parameters, but it would be much slower than using the earlier introduced FOURIER randomization.

Time Complexity of the Randomizations. The time complexity of the UNIFORM, INTERVAL and SUBSEQUENCE randomizations is $O(N)$, and of FOURIER $O(N \log N)$. MCMC DISTANCE randomization is in practice always much slower, but its time complexity cannot be given for a general case because the number

of MCMC iterations needed depends on the original sequence and the distance function. The time required by one MCMC iteration is typically dominated by the time complexity of the distance function. However, typical wall-clock running times to produce 1000 MCMC samples from the word and interbeat interval datasets used here are on the order of 1–3 and 10–20 minutes, respectively, using a non-optimized R [27] implementation and a standard desktop computer.

2.3 Properties of Fourier Parameters

In this section we show some properties of the Fourier parameters under the INTERVAL distribution: (i) the Fourier amplitudes are uncorrelated and their variance is proportional to σ_z^2, (ii) the phases φ_k approximately obey the uniform distribution on $[0, 2\pi)$.

Theorem 1. *The Fourier parameters satisfy the following properties under the* INTERVAL *distribution:*

- *The coefficient a_0 is the mean of the intervals in S_z.*
- *The expectations $E(a_k)$ and $E(b_k)$ vanish for every k.*
- *The variances are $E(a_k a_l) = E(b_k b_l) = 2\delta_{kl}\sigma_z^2(N-2)/(N(N-1))$ for every k and l, where δ_{kl} is the Kronecker delta.*
- *The cross-correlations $E(a_k b_l)$ vanish for every k and l.*
- *For all k, the phases $\varphi_k + 2\pi l/N \pmod{2\pi}$ are equally probable for all values of $l \in \{0, \ldots, N-1\}$.*

We omit the proof for brevity.

2.4 MCMC and Parallel Tempering

We use MCMC integration with parallel tempering [9,22] to draw samples from the distribution f defined by Equation (3). Instead of drawing samples directly from f the samples are drawn from a product distribution F

$$F(\{S'_\alpha\}_{\alpha\in\Lambda}) \propto \prod_{\alpha\in\Lambda} f(S'_\alpha)^\alpha, \tag{4}$$

where $\{0, 1\} \subseteq \Lambda \subseteq [0, 1]$ is a finite set and S'_α is a permutation of the original sequence. At each MCMC iteration, the value of S'_1 gives a sample from f.

The distribution for specific values of α in Λ are called "chains". We perform sampling using the Metropolis-Hastings algorithm in which the proposal distributions include changes into one chain (within-chain jumps) and swapping chains with adjacent values of α (chain swaps). Here we use the following proposal distributions for within-chain jumps, repeated 10 times per MCMC iteration: (i) permuting the interval sequence in random, (ii) permuting a randomly chosen subsequence of the sequence in random, (iii) reversing a randomly chosen subsequence, (iv) swapping randomly chosen intervals, and (v) swapping adjacent items.

The idea is that the chain mixes well for low values of α ("high temperatures"). Indeed, for the chain with $\alpha = 0$ each consecutive state is a random permutation of the interval sequence. If the set Λ is chosen suitably then there is a sufficient number of chain swaps, which brings states from the well-mixed high temperature chains into the target distribution for which $\alpha = 1$; we have verified that the chains mix sufficiently with the high temperatures and hence include a sufficient number of practically independent samples for use in the computation of confidence intervals. See [22] for a more detailed discussion on parallel tempering.

2.5 Confidence Intervals and Hypothesis Testing

Confidence intervals cannot in general be determined analytically for non-trivial features of interest and must hence be obtained by simulation. Confidence intervals of level α for any feature of interest (e.g., Fourier amplitudes, Fourier phases or the autocorrelation structure) can be computed by calculating the value of the feature for a set of simulated samples obtained from a chosen null distribution. The confidence intervals here are defined to be the $\alpha/2$ and $1-\alpha/2$ quantiles, where we always use $\alpha = 0.05$ in confidence levels and as a limit of significance. The parameters of interest can also be averaged in bins, in which case the confidence intervals will be narrower.[1] In this paper, we use the term *significant feature* to denote features that are outside the confidence intervals of some chosen null distribution. We further define a *non-random feature* as a feature that lies outside the confidence intervals calculated using the INTERVAL distribution.

The Fourier phases are approximately uniformly distributed under INTERVAL randomization (see Theorem 1). Due to the cyclic nature of the phases, confidence intervals for the phases cannot be computed in a meaningful way. Instead, the null hypothesis that the phases are uniformly distributed on the interval $[0, 2\pi)$ is tested by the Kolmogorov-Smirnov test.

3 Experiments

3.1 Datasets

The application of the randomization methods presented in this paper are illustrated using one artificial and two real-life datasets.

Toy Dataset. The toy dataset consists of two sequences: **(1)** The `AR` sequence is an autoregressive sequence of order 1 obeying $z_{n+1} \sim N(z_n, \varepsilon)$ and **(2)** the `periodic` sequence obeying $z_n \sim N(\cos(2\pi k_{\text{toy}} n/N), \varepsilon)$ is a cosine embedded in noise. $N(\mu, \sigma)$ denotes a normal distribution with mean μ and standard deviation σ; here we have used $k_{\text{toy}} = 7$ and $\varepsilon = 0.7$.

IBI Dataset. The signals in the IBI dataset are interval sequences representing the time between two successive heartbeats, forming an interbeat-interval (IBI)

[1] In this paper, we always use bins of width one for the word data and bins of width 10 for the heartbeat data.

series. It has been shown that the IBI series have different time domain (e.g. [23]) and frequency domain (e.g. [1]) properties for normal subjects and for subjects with heart failure. The IBI dataset was hence formed from two different datasets from the PhysioBank biomedical signal archive [11]: **(1)** The `normal rhythm` dataset[2] contains recordings from 54 subjects in normal sinus rhythm, **(2)** The `heart failure` dataset[3]) contains recordings from 29 subjects with congestive heart failure. The first 1500 intervals were chosen for analysis, which translates to about 25 minutes of data at a heart rate of 60 beats per minute.

Word Dataset. The word dataset is composed of words from the book *Pride and Prejudice* by Jane Austen, publicly available from Project Gutenberg[4]. The interval sequence in this case represents the number of words between successive occurrences of a particular word.[5] A previously reported [19] representative set of words was chosen for analysis, forming the **(1)** `bursty` and **(2)** `non-bursty` datasets, each containing 12 words. In addition, the words were also divided into frequency classes of low, medium and high, corresponding to frequencies of roughly 40, 200 and 1200, respectively.

3.2 Null Model Selection and Confidence Intervals

Here we demonstrate the advantages of the INTERVAL distribution over the UNIFORM distribution. In Fig. 1, both UNIFORM and INTERVAL randomizations are shown for the word *met*. We notice that for the UNIFORM randomization many of the Fourier amplitudes are significant, whereas for the INTERVAL randomization all Fourier parameters are consistent with the random data. This shows that there is structure in *met* not present in the UNIFORM distribution, but explained by the INTERVAL distribution. Observing non-random features under the INTERVAL distribution is always due to the ordering of the intervals, which is not the case if the UNIFORM distribution is used, as shown by the example in Fig. 1. See [21] for further discussion on the unsuitability of the uniform distribution in the analysis of interval sequences.

3.3 Investigating the Structure of the Datasets

In this section, we present an overview of the datasets and illustrate their general properties using examples. We determine non-random features of the sequences by calculating confidence intervals in accordance with Section 3.2, using the INTERVAL distribution as the null model.

The Toy Dataset. The Toy dataset is shown in Fig. 2. For `AR`, most of the Fourier coefficients are non-random, as the complex structure of the sequence

[2] `http://www.physionet.org/physiobank/database/nsr2db/`
[3] `http://www.physionet.org/physiobank/database/chf2db/`
[4] `http://www.gutenberg.org`
[5] More specifically, the interval is one plus the number of words between successive occurrences of a word, i.e., adjacent words have an interval value of one.

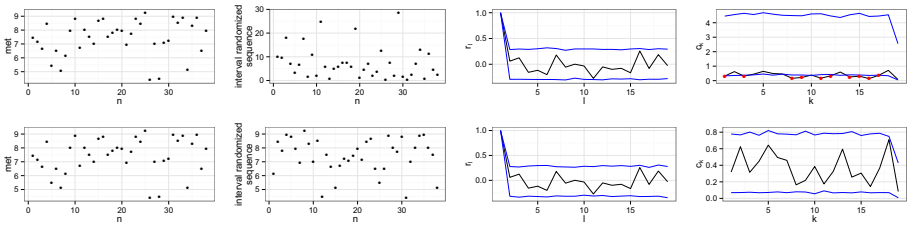


Fig. 1. UNIFORM (top) and INTERVAL (bottom) randomizations of the word *met*. Shown are, from the left: (1) the original sequence, (2) a realization from the random distribution, (3) the autocorrelation function, and (4) the Fourier amplitudes. Confidence intervals are shown in blue. Values outside confidence intervals are shown in red.

cannot be easily captured by a low number of Fourier amplitudes. For `periodic`, the Fourier coefficient for $k = 7$ is clearly non-random, corresponding to the number of periods in the sequence. The other Fourier amplitudes are mostly non-random, except for some high frequencies corresponding to the noise. INTERVAL randomization does not explain the autocorrelation structure of either sequence.

The IBI Dataset. Example sequences from the IBI dataset are shown in Fig. 3. Sequences from both the `normal rhythm` and `heart failure` datasets exhibit clear temporal structures, caused e.g. by different activities undertaken by the subject. This leads to segments with varying IBI distributions within one record. The temporal structure of the IBI sequences is usually characterized by a slow global trend containing segments with more rapid local variation.

The Fourier amplitudes for records with a strong temporal structure are only partially explained under the INTERVAL distribution. For such records, most of the Fourier amplitudes are non-random (see records `chf201` and `nsr033` in Fig. 3).

The non-random low-order Fourier amplitudes probably reflect the global trend, whereas the higher-order non-random Fourier amplitudes likely reflect short-range temporal variation. In contrast, some records with a weak global

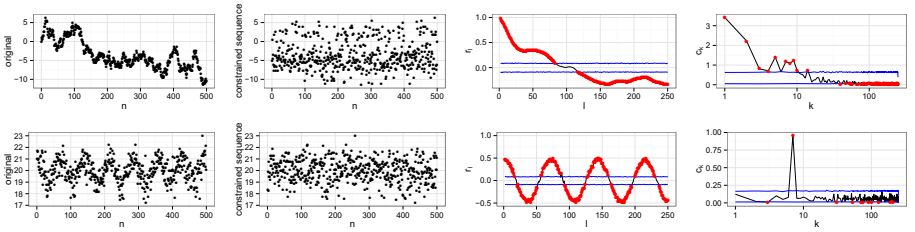


Fig. 2. The sequences in the Toy dataset. The `AR` (top row) and the `periodic` sequences (bottom row). Subplots follow the same order as in Fig. 1. The confidence intervals are based on INTERVAL randomization.

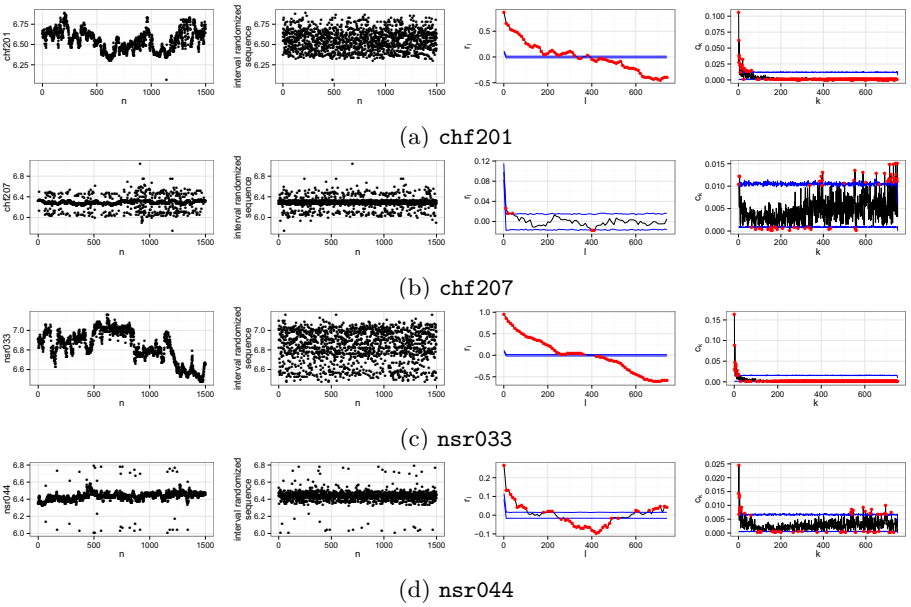


Fig. 3. Example IBI sequences. Subfigures and confidence intervals as in Fig. 2.

trend (`nsr044` in Fig. 3) or a high degree of outliers (`chf207` in Fig. 3) are better explained by the INTERVAL distribution, and for these records only a few Fourier coefficients are non-random.

The significant temporal structure of the sequences is also strongly reflected in the autocorrelation, which for most records is non-random.

The Word Dataset. Three example sequences from the word dataset are shown in Figs. 1 and 4. Only a few Fourier amplitudes or autocorrelation lags are non-random, usually marking visible temporal patterns in the data. The words *met* (Fig. 1) and *soon* (Fig. 4) are explained by the INTERVAL distribution. The word

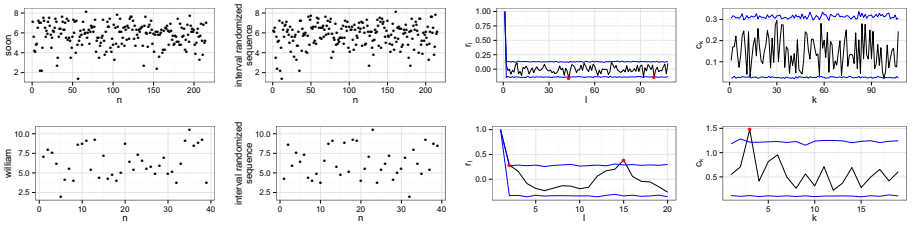


Fig. 4. Examples of word sequences. *Soon* (top) is a frequent word mostly (but not completely) explained by the INTERVAL distribution, *William* (bottom) contains a clear temporal structure. Subfigures and confidence intervals as in Fig. 2.

William (Fig. 4) contains a temporal structure that does not fit the INTERVAL distribution. The confidence intervals for the Fourier amplitudes are wider for low frequency words (mean confidence interval width 0.9) than for medium (0.3) and low (0.1) frequency words. Also, within a frequency class, confidence intervals are wider for `bursty` than for `non-bursty` words, the difference being 0.6 for low frequency, 0.1 for medium and 0.05 for high frequency words. Both observations are explained by the variance of the sequence, which follows a similar pattern (see Theorem 1).

Significant Features in the Datasets. The proportions of Fourier amplitudes, Fourier phases, and autocorrelation lags not explained by the INTERVAL randomization are shown in Tab. 1 (left column). On average, well over half of these features are non-random for the IBI sequences, compared to only a few percent for the word sequences. Therefore, the word dataset is better explained by the INTERVAL randomization than the IBI data. Furthermore, records in the `heart failure` dataset are generally better explained by the INTERVAL distribution than records in the `normal rhythm` dataset. This is likely due to the greater amount of outlier beats in the `heart failure` dataset and weaker global trends. The Fourier phases of the IBI data contain some non-random structure, but all phases in the word sequences are uniformly distributed.

3.4 Constrained Randomizations

We construct constrained randomizations by fixing a specific set of features in the FOURIER or DISTANCE randomizations. If the data are explained by the constrained null hypothesis, we can conclude that we have successfully located the features explaining the non-random characteristics of the data.

Connection between Fourier Amplitudes and Autocorrelation. In this section, both the `AR` and `periodic` sequences are randomized by fixing the most significant feature (with respect to INTERVAL randomization, see Fig. 2). The fixing of features is performed separately for autocorrelations and Fourier amplitudes. The results are shown in Fig. 5.

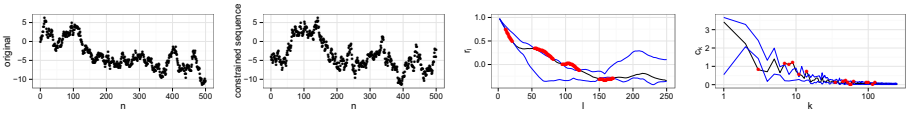
Table 1. Percentages of non-random features in the different datasets. The values represent mean (standard error of the mean) for c_k and r_l, and the percentage of sequences with non-uniform phases for φ_k (see Section 2.5).

		INTERVAL			FOURIER (c_k)			DISTANCE (r_l)		
Dataset		c_k	r_l	φ_k	c_k	r_l	φ_k	c_k	r_l	φ_k
IBI	`normal rhythm`	89 (1.5)	94 (0.9)	69	*0.2 (0.07)*	2 (0.5)	0	5 (0.6)	*3 (0.4)*	5
	`heart failure`	66 (4.7)	81 (4.5)	30	*0.8 (0.2)*	6 (1.3)	0	7 (1.7)	*2 (0.4)*	5
Word	`bursty`	6 (1.0)	10 (1.7)	0	*0.4 (0.2)*	2 (0.4)	0	9 (2.6)	*8 (2)*	6
	`non-bursty`	4 (0.9)	3 (0.9)	0	*0.5 (0.1)*	1 (0.4)	0	8 (0.9)	*5 (0.9)*	4

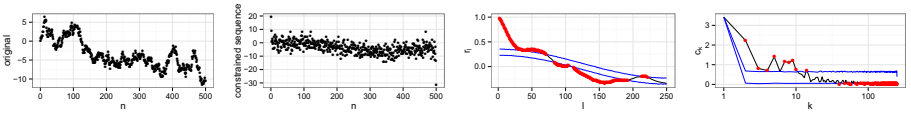
For `AR`, fixing the autocorrelation r_1 produces a sequence that retains the local temporal structure of the original sequence. The Fourier amplitudes are almost explained, but the autocorrelation function matches the original only for short lags. In contrast, fixing a single non-random feature in the Fourier domain performs much worse in explaining the data.

For `periodic`, Fourier amplitude randomization yields signals resembling the original. The majority of the Fourier amplitudes are explained, and the confidence intervals for the autocorrelations follow the course of the original autocorrelation function, albeit not perfectly. In contrast, fixing a single autocorrelation lag for `periodic` does not explain the features of the signal at all.

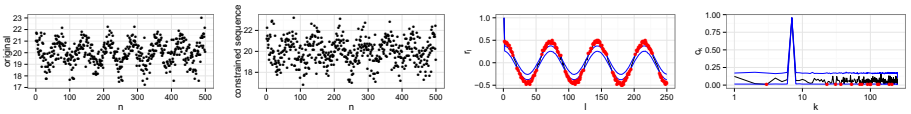
Fixing Fourier Amplitudes and Autocorrelation Structure. Constrained randomization of Fourier amplitudes and autocorrelation lags was applied to both the IBI and word datasets, keeping the non-random Fourier amplitudes and autocorrelation lags constant. The percentage of Fourier amplitudes and autocorrelation lags that remain significant under the randomizations are shown in Tab. 1. For the constrained Fourier amplitude randomization (middle column) the percentages are low, indicating that the data are well explained. Only the autocorrelations of the `heart failure` dataset shows a slightly higher percentage of significant features. For the constrained autocorrelation randomization



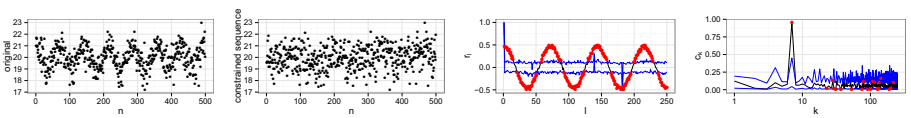
(a) The `AR` sequence randomized by fixing r_1.



(b) The `AR` sequence randomized by fixing c_1.



(c) The `periodic` sequence randomized by fixing c_7.



(d) The `periodic` sequence randomized by fixing r_{182}.

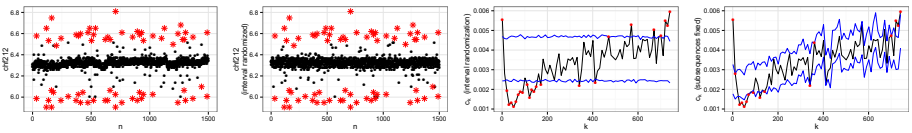
Fig. 5. Constrained randomizations of the toy data. Subplots are as in Fig. 1.

(rightmost column), the Fourier amplitudes of the word data and `heart failure` IBI data have above 5% of significant features, indicating that the randomization does not fully explain the Fourier amplitudes. There is also unexplained autocorrelation structure in the word data.

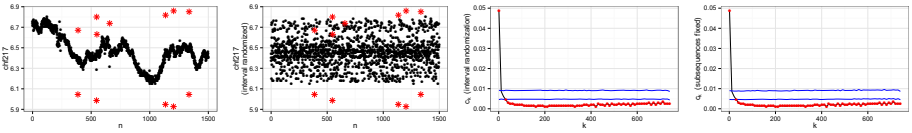
Since 95% confidence intervals were used, one must note in the interpretation of Tab. 1, that if the randomization explains the data, at most 5% of features should remain significant. In practice, this value is lower as a large portion of the features is kept fixed, especially for the IBI data. Also, the autocorrelation lags are not independent, causing them to fit the simple quantile based confidence intervals better than Fourier amplitudes.

Fixed Subsequence Randomization. Outliers in the data can significantly affect the structure and interpretation of the data. In order to investigate the structure of the data, outliers can be considered subsequences and kept fixed in the SUBSEQUENCE randomization.

In Fig. 6, outliers detected using a commonly used algorithm by [37] were kept fixed while the rest of the data were randomized using the SUBSEQUENCE method. In Fig. 6a several of the Fourier coefficients are outside the confidence intervals calculated using INTERVAL randomization, i.e., the structure of the data is not modeled by the INTERVAL distribution. However, fixing the outliers and calculating the confidence intervals using the SUBSEQUENCE distribution explains the data. In contrast, the Fourier amplitudes in Fig. 6b remain outside the confidence intervals even after fixing the outliers. This indicates that a more sophisticated method should be used to explain the remaining structure.



(a) `chf212`



(b) `chf217`

Fig. 6. Application of fixed subsequence randomization. Outliers in the sequences (plotted as red stars) are kept fixed during INTERVAL randomization. The plots show (1) the original data, (2) a realization of INTERVAL randomized data, (3) the original Fourier coefficients and INTERVAL confidence intervals and (4) the original Fourier coefficients with SUBSEQUENCE confidence intervals.

3.5 Application to Hypothesis Testing

Constrained realizations obtained e.g. by fixing non-random Fourier amplitudes can be used in statistical hypothesis testing. As an example of this, the significance of the pNN50-value commonly used in heart rate variability analysis [23] was calculated[6] for the records in the IBI datasets. The results are shown in Tab. 2. There are clearly differences between the choices of constraints.

On one hand, the INTERVAL randomization appears to provide realizations that are consistently too extreme for hypothesis testing, at least if the objective is to study the differences between normal rhythm and heart failure. On the other hand, the number of significant p-values with the Fourier amplitude constraint is much smaller for `normal rhythm` than for `heart failure`, suggesting that part of the IBI signal measured by the pNN50 statistic and not explained by the amplitudes, is related to the heart failure condition. Therefore, the null hypothesis with the amplitude constraint might be suitable for modeling healthy individuals.

Table 2. Percentages (%) of significant pNN50 -values for the datasets. From the left: by constraining Fourier coefficients, phases, autocorrelation lags, and using INTERVAL randomization.

	Randomization method			
Dataset	FOURIER (c_k)	FOURIER (φ_k)	DISTANCE (r_l)	INTERVAL
`normal rhythm`	22.2	75.9	100.0	100.0
`heart failure`	58.6	79.3	86.2	96.5

4 Related Work

Randomization testing in statistical analysis has a long history; see, e.g., [12,36] for a review. Randomization methods are useful in hypothesis testing and defining confidence bounds when sampling from the null hypothesis is easier than to define the null hypothesis analytically. Randomization methods have been devised for various kinds of data structures, such as binary matrices [10], graphs [13,38], gene periodicity (e.g., [15]), and real matrices [24].

Time series randomization has been studied, e.g., in [3,16,25,34,2,30,35]. Some of the prior randomization methods work in the Fourier space (see, e.g., [26] for use of phase-randomization in hypothesis testing) or in the wavelet space, see [17] for a review. However, usually the time series has not been represented as (equally-spaced) sequence of intervals, but as an event sequence with variable event interval (see, e.g., [4,31]).

In the field of data analysis, a recently promoted approach [14,20,5,6] to the use of randomization is to interpret the patterns as constraints to the null hypothesis. The use of surrogate data [32] in the hypothesis testing of data

[6] Here we consider all interbeat intervals, not just normal-to-normal (NN) intervals.

structures is a common technique, and has been applied in the generation of constrained realizations for hypothesis testing regarding the properties of a time series, e.g., by [33,29,28].

Randomization techniques have been applied in the analysis of heart rate variability (HRV), e.g., by [18,8], who used Fourier phase randomization for generating surrogate data for hypothesis testing. Time-varying surrogates were used by [7] for studying non-linearity in interbeat interval (IBI) series.

5 Conclusions

We have shown that interval sequences form a natural representation for event sequences, and offer a principled and robust basis to sequence randomization. We have investigated the problem of interpreting commonly used Fourier parameters and autocorrelation structures. We find that the interpretation depends on the null hypothesis used; for example, a naïve use of the UNIFORM distribution may lead to false conclusions regarding the temporal structure of sequences.

Furthermore, we have provided computationally efficient randomization methods for studying Fourier parameters, and an MCMC based method for studying autocorrelation structures and arbitrary constraints. The randomization methods allow the user to efficiently test different null hypotheses by fixing chosen subsets of parameters. This makes it possible to infer possible causes for the observed significant patterns.

In this paper, we have shown how the proposed randomization methods can be used in hypothesis testing, and examined the role of the null hypothesis. There is no universally suitable null hypothesis. The null hypothesis should encompass our best understanding of the features of the data and hence depends on the research question.

With the help of the randomization methods presented here, simple and understandable explanations for the structure of the data can be found efficiently and in a statistically robust way. If there are structures left unexplained by the proposed methods, more complex constraints or models of different types can be used to further investigate and explain the remaining patterns in the data.

Acknowledgements. This study was supported by the SalWe Research Programme for Mind and Body (Tekes – the Finnish Funding Agency for Technology and Innovation grant 1104/10).

References

1. Bigger, J.T., Fleiss, J.L., Steinman, R.C., Rolnitzky, L.M., Schneider, W.J., Stein, P.K.: RR variability in healthy, middle-aged persons compared with patients with chronic coronary heart disease or recent acute myocardial infarction. Circulation 91(7), 1936–1943 (1995)

2. Bullmore, E., Long, C., Suckling, J., Fadili, J., Calvert, G., Zelaya, F., Carpenter, T.A., Brammer, M.: Colored noise and computational inference in neurophysiological (fMRI) time series analysis: Resampling methods in time and wavelet domains. Human Brain Mapping 12(2), 61–78 (2001)
3. Carlstein, E.G.: Resampling techniques for stationary time-series: some recent developments. University of North Carolina at Chapel Hill (1990)
4. Clifford, G.D., Azuaje, F., McSharry, P., et al. (eds.): Advanced Methods and Tools for ECG Data Analysis. Artech House, London (2006)
5. De Bie, T.: An information theoretic framework for data mining. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2011, pp. 564–572. ACM, New York (2011)
6. De Bie, T.: Maximum entropy models and subjective interestingness: an application to tiles in binary databases. Data Mining and Knowledge Discovery 23(3), 407–446 (2011)
7. Faes, L., Zhao, H., Chon, K., Nollo, G.: Time-varying surrogate data to assess nonlinearity in nonstationary time series: Application to heart rate variability. IEEE Transactions on Biomedical Engineering 56(3), 685–695 (2009)
8. Garde, S., Regalado, M.G., Schechtman, V.L., Khoo, M.C.: Nonlinear dynamics of heart rate variability in cocaine-exposed neonates during sleep. American Journal of Physiology-Heart and Circulatory Physiology 280(6), H2920–H2928 (2001)
9. Geyer, C.J.: Markov chain Monte Carlo Maximum Likelihood. In: Computing Science and Statistics: The 23rd Symposium on the Interface, pp. 156–163. Interface Foundation, Fairfax (1991)
10. Gionis, A., Mannila, H., Mielikäinen, T., Tsaparas, P.: Assessing data mining results via swap randomization. ACM Trans. Knowl. Discov. Data 1(3) (December 2007)
11. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E.: PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 101(23), e215–e220 (2000)
12. Good, P.I.: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. Springer (2000)
13. Hanhijärvi, S., Garriga, G.C., Puolamäki, K.: Randomization techniques for graphs. In: Proceedings of the 9th SIAM International Conference on Data Mining (SDM 2009), pp. 780–791 (2009)
14. Hanhijärvi, S., Ojala, M., Vuokko, N., Puolamäki, K., Tatti, N., Mannila, H.: Tell me something I don't know: randomization strategies for iterative data mining. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2009, pp. 379–388. ACM, New York (2009)
15. Kallio, A., Vuokko, N., Ojala, M., Haiminen, N., Mannila, H.: Randomization techniques for assessing the significance of gene periodicity results. BMC Bioinformatics 12(1), 330 (2011)
16. Kreiss, J.P., Franke, J.: Bootstrapping stationary autoregressive moving-average models. Journal of Time Series Analysis 13(4), 297–317 (1992)
17. Laird, A.R., Rogers, B.P., Meyerand, M.E.: Comparison of fourier and wavelet resampling methods. Magnetic Resonance in Medicine 51(2), 418–422 (2004)
18. Li, C., Ding, G.H., Wu, G.Q., Poon, C.S.: Band-phase-randomized surrogate data reveal high-frequency chaos in heart rate variability. In: 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 2806–2809 (2010)

19. Lijffijt, J., Papapetrou, P., Puolamäki, K.: Size matters: Finding the most informative set of window lengths. In: Flach, P.A., De Bie, T., Cristianini, N. (eds.) ECML PKDD 2012, Part II. LNCS, vol. 7524, pp. 451–466. Springer, Heidelberg (2012)
20. Lijffijt, J., Papapetrou, P., Puolamäki, K.: A statistical significance testing approach to mining the most informative set of patterns. Data Mining and Knowledge Discovery (December 2012) (to appear) (published online before print)
21. Lijffijt, J., Papapetrou, P., Puolamäki, K., Mannila, H.: Analyzing word frequencies in large text corpora using inter-arrival times and bootstrapping. In: Gunopulos, D., Hofmann, T., Malerba, D., Vazirgiannis, M. (eds.) ECML PKDD 2011, Part II. LNCS, vol. 6912, pp. 341–357. Springer, Heidelberg (2011)
22. Liu, J.: Monte Carlo Strategies in Scientific Computing. Series in Statistics. Springer (2008)
23. Mietus, J., Peng, C., Henry, I., Goldsmith, R., Goldberger, A.: The pnnx files: re-examining a widely used heart rate variability measure. Heart 88(4), 378–380 (2002)
24. Ojala, M., Vuokko, N., Kallio, A., Haiminen, N., Mannila, H.: Randomization methods for assessing data analysis results on real-valued matrices. Statistical Analysis and Data Mining 2(4), 209–230 (2009)
25. Politis, D.N.: The impact of bootstrap methods on time series analysis. Statistical Science 18(2), 219–230 (2003)
26. Prichard, D., Theiler, J.: Generating surrogate data for time series with several simultaneously masured variables. Physical Review Letters 73(7), 951–954 (1994)
27. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013) ISBN 3-900051-07-0, http://www.R-project.org/
28. Schreiber, T.: Constrained randomization of time series data. Physical Review Letters 80(10), 2105–2108 (1998)
29. Schreiber, T., Schmitz, A.: Improved Surrogate Data for Nonlinearity Tests. Physical Review Letters 77(4), 635–638 (1996)
30. Schreiber, T., Schmitz, A.: Surrogate time series. Physica D: Nonlinear Phenomena 142(3-4), 346–382 (2000)
31. Sörnmo, L., Laguna, P.: Bioelectrical Signal Processing in Cardiac and Neurological Applications. Academic Press (2005)
32. Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Doyne Farmer, J.: Testing for nonlinearity in time series: the method of surrogate data. Physica D: Nonlinear Phenomena 58(1), 77–94 (1992)
33. Theiler, J., Prichard, D.: Constrained-realization Monte-Carlo method for hypothesis testing. Physica D: Nonlinear Phenomena 94(4), 221–235 (1996)
34. Vinod, H.D.: Maximum entropy ensembles for time series inference in economics. Journal of Asian Economics 17(6), 955–978 (2006)
35. Vuokko, N., Kaski, P.: Significance of patterns in time series collections. In: Proceedings of the Eleventh SIAM International Conference on Data Mining, Mesa, AZ, April 28-30, pp. 676–686. SIAM, Philadelphia (2011)
36. Westfall, P.H., Young, S.: Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment. A Wiley-Interscience publication, Wiley (1993)
37. Xu, X., Schuckers, S.: Automatic detection of artifacts in heart period data. Journal of Electrocardiology 34(4), 205–210 (2001)
38. Ying, X., Wu, X.: Graph generation with prescribed feature constraints. In: Proceedings of the 9th SIAM International Conference on Data Mining (SDM 2009), pp. 966–977 (2009)