PATTERN DISCOVERY IN TIME SERIES
A SURVEY

DENIS KHRYASHCHEV, GRADUATE CENTER, CUNY, OCTOBER 2018
Often datasets represent processes that take place over long periods of time. Their outputs are measured at regular time intervals creating discrete time series. For example, consider CitiBike demand and Fisher river temperature data.

**CitiBike**

**Fisher river**

MOTIVATION. COMPLEXITY

Complexity quantifies the internal structure of the underlying process. EEG data can be classified [1] into interictal, preictal and seizure using their complexity.

MOTIVATION. PERIODICITY

Natural phenomena like Sun activity, Earth rotation and revolution drive periodic human activity on the large scale. E.g. New York City’s human mobility is highly periodic with clear peaks in ridership from 6 AM to 10 AM, and from 3 PM to 7 PM.

Predictability estimates the expected accuracy of forecasting given time series. Often there is a trade-off between the desired accuracy and computation time [2].

MOTIVATION. CLUSTERING

Often a task of grouping similar in certain quality time series arises in the domains of transportation, finance, medicine,… Time sensitive modifications of standard techniques are applied, e.g. k-means of autocorrelation functions.

Image source: Denis Khryashchev’s summer internship at Simulmedia (Jun – Aug 2018).
Perhaps, the most well known and widely applied task related to time series is forecasting. Understanding time series periodicity, complexity, and predictability helps in selecting better predictors and optimizing parameters. E.g., knowing periodicity \( P=5 \) of the series, one can forecast averaging values with lag 5.

Video source: Denis Khryashchev’s summer internship at Simulmedia (Jun – Aug 2018).
Throughout the presentation we will consider time series of real values and will use the following notation

\[ X = \{X_1, \ldots, X_N\} = \{X_t\}_{t=1}^{N}, X_t \in \mathbb{R} \]

Not to be confused with set notation, \( \{\cdot\} \) is used to denote sequences.

A subsequence of the series \( X \) that starts at period \( i \) and ends at period \( j \) is written as

\[ X_i^j = \{X_t\}_{i}^{j} = \{X_i, \ldots, X_j\}, i \leq j \]
ORGANIZATION OF THE PRESENTATION

1. Complexity
2. Entropy
3. Predictability
4. Periodicity
5. Similarity
6. Clustering
7. Forecasting
1. KOLMOGOROV COMPLEXITY

For time series $X$ we define the Kolmogorov complexity as the length of the shortest description of a sequence of values ordered in time in some fixed universal description language

$$K(X) = |d(X)|$$

where $K$ is the Kolmogorov complexity, and $d(X)$ is the shortest description of the time series $X$. Smaller values of $K(X)$ indicate lower complexity.
Given two time series

\[ X = \{0, 1, 0, 1, 0, 1, 0, 1\} \] and \[ Y = \{1, 0, 0, 1, 1, 1, 0, 0, 1, 0\} \]

and selecting Python as our description language we have the shortest descriptions

\[ d_P(X) = [0, 1] \ast 5 \] and \[ d_P(Y) = [1, 0, 0, 1, 1, 1, 0, 0, 1, 0] \]

quantifying smaller “Pythonic” complexity for \( X \) comparing to \( Y \)

\[ K_P(X) = |d_P(X)| = 7 \]
\[ K_P(Y) = |d_P(Y)| = 21 \]
However, as proven by Kolmogorov in [3], and Chaitin and Arslanov in [4], the complexity $K$ is not a computable function in general.

1. LEMPEL-ZIV COMPLEXITY

Lempel and Ziv [5] proposed a combinatorial approximation of the complexity of finite sequences based on their production history. For time series $X$ it is

$$H(X) = \{X_1^{h_1+1}, X_{h_1+1}^{h_2}, \ldots, X_{h_{m-1}+1}^m\}$$

For series $X = \{0,0,0,1,1,0,1,0,0,1,0,0,1,0,1\}$ one of the production histories is

$$H(X) = \{0\} \cup \{0,0,1\} \cup \{1,0\} \cup \{1,0,0\} \cup \{1,0,0,0\} \cup \{1,0,1\}$$

The overall complexity is the size of the shortest possible production history

$$c(X) = \min_{H(X)} |H(X)|$$

**Disadvantage:** the actual values $X_t$ are treated as symbols, e.g.

$$c(X = \{1, 2, 1, 5, 1, 2\}) = c(Y = \{8, 0.5, 8, 0.1, 8, 0.5\})$$

2. ENTROPY

Shannon and Weaver introduced entropy [6] as a measure of information transmitted by a signal in a communication channel

\[ H(X) = -\mathbb{E}[\log_2 P(X)] \]

Renyi [7] generalized the definition for ordinary discrete finite distribution of \( X \)
\( \mathcal{P} = (p_1, ..., p_M), \sum_k p_k = 1 \) to entropy of order \( \alpha \) (\( \alpha \to 1 \) for Shannon entropy)

\[ H_\alpha(X) = H_\alpha(\mathcal{P}) = \frac{1}{1-\alpha} \log_2 (\sum_k p_k^\alpha) \]

**Disadvantage:** both definitions do not take order of the values \( X_t \) into account, e.g. \( H(X = \{1, 2, 3, 1, 2, 3\}) = H(Y = \{1, 3, 2, 2, 3, 1\}) \).

2. KOLMOGOROV ENTROPY

Entropy is often used as an approximation of complexity. Among the most well-known approximations [8] of the complexity is Kolmogorov Entropy defined as

\[ K = - \lim_{\tau \to \infty} \lim_{\epsilon \to \infty} \lim_{d \to \infty} \frac{1}{d \tau} \sum_{i_1, \ldots, i_d} p(i_1, \ldots, i_d) \ln p(i_1, \ldots, i_d) \]

It describes complexity of a dynamic system with \( F \) degrees of freedom. \( F \)-dimensional phase space is partitioned into \( \epsilon^F \) boxes, \( \tau \) stands for time intervals, and \( p(i_1, \ldots, i_d) \) is the joint probability that we find the \( F \)-dimensional point representing values \( X_{t=k\tau} \) inside the box \( \epsilon^F \).

**Disadvantage:** the approximation is computable for known analytically defined models, however, it is hard to calculate it given the resulting series only.

2. ENTROPY WITH TEMPORAL COMPONENT

Another definition [6] of entropy takes into account temporal order of the values $X_t$

$$H_t(X) = - \sum_{i=1}^{N} \sum_{j=1}^{N} P(X^j_i) \log_2(P(X^j_i))$$

$P(X^j_i)$ is the probability of the subsequence $X^j_i$. $H_t(X)$ is $O(2^N)$ complex. Lempel-Ziv estimator [9] approximates $H_t(X)$ rapidly converging

$$H_{LZ}(X) = \left(\frac{1}{N} \sum_t s'_t\right)^{-1} \ln N$$

where $s'_t$ is the shortest subsequence starting at period $t$ observed for the 1st time.

**Disadvantage:** values $X_t$ are treated as symbols, e.g.

$$H_{LZ}(X = \{1, 2, 1, 5\}) = H_{LZ}(Y = \{2, 9, 2, 3\})$$

Bandt and Pompe [10] proposed permutation entropy of order $n$

$$H(n) = - \sum p(\pi) \log p(\pi)$$

where $p(\pi) = \frac{\# \{ t | 0 \leq t \leq T-n, \text{type}(X_{t+1}, \ldots, X_{t+n}) = \pi \}}{T-n+1}$ is a frequency of a permutation of type $\pi$. E.g., for $X = \{4, 7, 9, 10, 6, 11, 3\}$, $n = 3$ we have

$\pi(4, 7, 9) = \pi(7, 9, 10) = \pi_{012}(X_t < X_{t+1} < X_{t+2})$,
$\pi(9, 10, 6) = \pi(6, 11, 3) = \pi_{210}(X_{t+2} < X_t < X_{t+1})$,
$\pi(10, 6, 11) = \pi_{102}(X_{t+1} < X_t < X_{t+2})$.

The entropy becomes $H(3) = -2 \frac{2}{5} \log \frac{2}{5} - \frac{1}{5} \log \frac{1}{5} \approx 1.52$.

**Disadvantage:** the definition requires $X_t \neq X_{t+1}$ and has a complexity of $O(n!)$. 

---

3. PREDICTABILITY

Following the Fano inequality [11], predictability of series $X \Pi(X) \leq \Pi^{max}(H[X], N)$ where $N$ is the number of unique values $X_t$, $0 \leq \Pi^{max} \leq 1$ is the maximum predictability of $X$ with 0 standing for completely unpredictable chaotic series.

In the works of Song et al. [12] the maximum predictability is shown to be the solution of

$$H(X) = -\Pi^{max} \log_2 \Pi^{max} - (1 - \Pi^{max}) \times \log_2 (1 - \Pi^{max}) + (1 - \Pi^{max}) \log_2 (N - 1).$$

Often Lempel-Ziv estimator $H_{LZ}(X)$ is used to calculate the entropy of the series.

**Disadvantage:** depends of the selected measure of $H(X)$, no closed-form solution of the equation.

3. PREDICTABILITY. SHUFFLING

Alternative approach to measure predictability was proposed by Kaboudan [13]. They defined it as a ratio of forecasting errors prior and after shuffling of the series

$$\eta(X) = 1 - \frac{(X-f(X))^T(X-f(X))}{(X^s-f(X^s))^T(X^s-f(X^s))}$$

where $X^s$ is a shuffled copy of the series, $f$ is the selected predictor.

**Disadvantage:** the shuffling is done only once and could lead to inconsistent results when measured multiple times for the same time series $X$. A significant improvement of the approach would be to calculate $\eta(X)$ numerous times (e.g. 1000) calculating $p$ value and performing hypothesis testing.

Often in econometrics predictability is quantified with linear regression models [14]

\[ X_{t+1} = \alpha + \beta X_t + \epsilon_{t+1} \]

that are used to calculate \( R^2 \)

\[ R^2 = 1 - \frac{\text{var}(\alpha + \beta X_t)}{\text{var}(X_{t+1})} \]

ranging from 0 to 1 with the latter representing the most unpredictable series.

**Disadvantage:** captures only linear relationships unless non-linear regression models are used.

4. PERIODICITY. FOURIER

Numerous approaches to quantify time series periodicity rely on Fourier transform

\[ M_k = \sum_{t=0}^{N-1} X_t e^{-i2\pi kt/N} \]

where \(M_k\) is the “magnitude” of \(k^{th}\) frequency quantifying “relative chance” of the repetition of the value \(X_t\) after the corresponding period of time.

To decrease the amount of spurious artifacts, usually a windowed or short-time Fourier transform is applied [15]

\[ M_k = \sum_{t=0}^{N-1} X_t w(t - \tau) e^{-i2\pi kt/N} \]

where \(w\) is the window function of effective size of \(2\tau\). Often Blackman, Hamming, and Bartlett windows are used.

**Disadvantage:** not linear in period.

Linear autocorrelation function can be used to evaluate periodicity due to the Wiener-Khinchin [16][17] theorem that states that

\[ S_{xx}(f) = \sum_{t=0}^{N-1} r_{xx} e^{-i2\pi kt/N} \]

where \( S_{xx}(f) \) is the power spectrum of \( X \), \( r_{xx} \) is its autocorrelation function. In other words, larger values of the autocorrelation function of \( X \) at lag \( \tau \) can signify larger periodicity of the series at lag \( \tau \).

4. PERIODICITY. FISHER TEST

A similar test based on the notion of periodograms was proposed by Fisher [18]

\[
I(f) = \frac{2}{N} \sum_{t=0}^{N-1} X_t \cos 2\pi ft + \frac{2}{N} \sum_{t=0}^{N-1} X_t \sin 2\pi ft
\]

\[
W = \max_f I(f)
\]

where frequency \(-\frac{1}{2} \leq f \leq \frac{1}{2}\). The test assumes that \(X = \zeta + a\) with \(\zeta\) being the real periodic signal and \(a \sim N(0, \sigma^2)\) is the unobserved noise. \(W\) is measured to reject the \(H_0: \zeta = 0\).

Disadvantage: similar to original Fourier transform – spurious artifacts. Can be improved with window functions.

4. PERIODICITY TRANSFORM

Sethares and Staley [19] proposed linear-in-period transformation of time series. They defined $X$ to be $p$-periodic if $X_t = X_{t+p}$ and $P_p$ to be a set of all $p$-periodic sequences. They introduced non-orthogonal basis sequences that are $p$-periodic

$$\delta_p^s(j) = \begin{cases} 1 & \text{if } (j - s) \mod p = 0 \\ 0 & \text{otherwise} \end{cases}$$

where $j$ is the time index of the series $X$, and $s$ is the time shift. The measure of periodicity is the projection

$$\pi(X, P_p) = \sum_{s=0}^{p-1} \alpha_s \delta_p^s, \quad \alpha_s = \frac{1}{N} \sum_{t=0}^{N-1} X_{s+tp}$$

Disadvantage: sample interval of $X$ or its entire length must correspond to an integer factor of $p$.

Linear correlation is a standard and well-known dependence measure. For two time series $X$ and $Y$ it is defined as

$$\rho_{X,Y} = \rho(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]}{\sigma_X \sigma_Y}$$

Often it is more convenient to standardize the series $X' = \frac{X - \mathbb{E}(X)}{\sigma_X}$ and $Y' = \frac{Y - \mathbb{E}(Y)}{\sigma_Y}$

$$\rho(X, Y) = \mathbb{E}[X'Y']$$

Ranging as $-1 \leq \rho(X, Y) \leq 1$ it captures linear dependence between $X$ and $Y$.

**Limitation:** does not capture non-linear relationship between $X$ and $Y$. 

**5. SIMILARITY AND DEPENDENCE. LINEAR CORRELATION**
Renyi considered a general measure of dependence $\delta(X,Y)$ and postulated:

- $\delta(X,Y)$ is defined for any pair of random variables $X,Y$ neither of them being constant with probability 1
- It is symmetric, $\delta(X,Y) = \delta(Y,X)$
- $\delta(X,Y) = 0$ if and only if $X$ and $Y$ are independent
- $0 \leq \delta(X,Y) \leq 1$
- $\delta(X,Y) = 1$ only if there is a strict dependence between $X$ and $Y$
- $\delta(X,Y) = \delta(\text{Id}(X), \text{Id}(Y))$ where $\text{Id}$ is a Borel-measurable identity function
- If the joint distribution of $X$ and $Y$ is normal then $\delta(X,Y) = |\rho(X,Y)|$. 

5. SIMILARITY AND DEPENDENCE.

MAXIMAL CORRELATION COEFFICIENT

The maximal correlation coefficient proposed by Gebelein [21] fits all of the Renyi’s postulates and is defined as

$$\rho_{\text{max}}(X, Y) = \max_{f, g} \rho(f(X), g(Y))$$

where \(f: \mathbb{R} \rightarrow \mathbb{R}, \ g: \mathbb{R} \rightarrow \mathbb{R}\) are Borel-measurable functions, \(0 \leq \text{Var}(f(X)) < B\), and \(0 \leq \text{Var}(g(Y)) < B, B \in \mathbb{R}\).

$$\rho_{\text{max}}(X, Y) = 0$$ if \(X\) and \(Y\) are independent.
$$\rho_{\text{max}}(X, Y) = 1$$ if either \(X = f(Y)\) or \(Y = g(X)\).

**Limitation:** The functions \(f^*, g^*\) that maximize \(\rho(f(X), g(Y))\) are not guaranteed to have inverses \((f^*)^{-1}, (g^*)^{-1}\).

5. MAXIMAL VS LINEAR AUTOCORRELATION

CitiBike pickup data has a daily periodicity depicted with linear autocorrelation function. However, maximal autocorrelation function demonstrates the presence of non-linear dependencies due to 
\[ \rho_{max}(X_1^{N-\tau}, X_\tau^N) > \rho(X_1^{N-\tau}, X_\tau^N) \] for every lag \( \tau \).
Dembo et al. [22] demonstrated that the maximal correlation coefficient equals to

$$\rho_{\text{max}}(S_m, S_n) = \sqrt{m/n}$$

where $S_k = \sum_{i=1}^{k} Y_k$, $Y_1, \ldots, Y_N$ is a collection of independent identically distributed random variables with $Var(Y_i) < \infty$.

Lancaster [23] has shown that if $(X, Y)$ is a bivariate Gaussian vector the maximal correlation coefficient is equal to

$$\rho_{\text{max}}(X, Y) = |\rho(X, Y)|$$

5. SOME PROPERTIES OF MAXIMAL CORRELATION

Witsenhausen [24] proposed a way to calculate the maximal correlation coefficient for discrete $X$ and $Y$:

- First, the ordered sets $\alpha$ and $\beta$ that contain unique values of $X$ and $Y$ are built.
- Second, the probabilities $p_{ij}$ from the contingency table for $X$ and $Y$ are calculated,
  
  \[ p_{ij} = P(X = \alpha_i | Y = \beta_j). \]

- Third, the normalized joint-probability matrix $Q = (q_{ij})$ is computed
  \[ q_{ij} = \frac{p_{ij}}{\sqrt{p(.i).p(.j)}}. \]

Then the maximal correlation coefficient is equal to

\[ \rho_{max}(X, Y) = \lambda_2 \]

where $\lambda_2$ is the second singular value of $Q$.

5. MONOTONE CORRELATION

Monotone correlation coefficient was proposed by Kimeldorf et al. [25] and was defined as follows: let $\mathcal{F} = \{f: \mathbb{R} \to \mathbb{R} | f \text{ is monotone}\}$ then the monotone correlation

$$\rho_{mono}(X, Y) = \max_{f, g \in \mathcal{F}} (f(X), g(Y))$$

Overall, we have the following relationship between linear, monotone and maximal correlation coefficients

$$0 \leq |\rho(X, Y)| \leq \rho_{mono}(X, Y) \leq \rho_{max}(X, Y) \leq 1$$

**Limitation:** there is no known formula that computes the value of the monotone correlation coefficient. It is a maximization problem.

6. CLUSTERING

We will denote $x$ as a collection of time series

$$x = \{X_1, ..., X_N\}, X_i = \{X_{i,1}, X_{i,2}, ..., X_{i,M}\}, X_{i,j} \in \mathbb{R}.$$ 

We define clustering as an unsupervised partitioning of $X$ into $k$ groups assigning 1 label from the set of labels $\{C_1, ..., C_k\}$ to every time series $X_i$ such that each label $C_i$ is assigned at least once.
6. CLUSTERING. NAÏVE APPROACH

Clustering discrete time series $X_1, X_2, X_3$ into two groups naïvely applying k-means results in clusters $\{X_2, X_3\}$ and $\{X_1\}$. However, it seems to be more reasonable to group $X_1$ and $X_2$ together. Using autocorrelation functions $R = \{r_{xx}(X_1), r_{xx}(X_2), r_{xx}(X_3)\}$ instead of the actual value will group $X_1$ and $X_2$ in the same cluster.

$$r_{xx}(X) = \{\rho(X, L^\tau X) | 2 \leq \tau \leq N - 1\}, L^k X_t = X_{t-\tau}$$
6. TYPES OF CLUSTERING APPROACHES

- Raw data. Standard clustering techniques applied to raw data with modified distance or dissimilarity measures.

- Feature generation. New features are generated from raw data and clustered with standard techniques.

- Model assumption. Clustering based on model parameters, hypothesis testing.
6. RAW DATA CLUSTERING

Komelj and Batagelj [26] modified relocation clustering introducing new similarity measure

\[
D(X,Y) = \sum_t \alpha_t d_t(X,Y), \alpha_s \neq \alpha_t, \alpha_t \geq 0, \sum_t \alpha_t = 1
\]

Golay et al. [27] proposed cross-correlation distance

\[
d_{cc}^1(X,Y) = \left( \frac{1-\rho(X,Y)}{1+\rho(X,Y)} \right)^\beta, \quad d_{cc}^2(X,Y) = 2\left(1 - \rho(X,Y)\right)
\]

**Limitations:** most of the approaches do not take into account sequences of values \(X_{t-k}, X_{t-k+1}, \ldots, X_t\) but instead compare neighboring values.

6. RAW DATA CLUSTERING

Moller-Levet et al. [28] described short time series distance

\[ d_{STS}^2(X, Y) = \sum_k \left( \frac{\delta X_k}{\delta t_k} - \frac{\delta Y_k}{\delta t_k} \right)^2, \delta X_k = X_k - X_{k-1} \]

Batista et. al [29] showed existing distance measures are not efficient for complex time series, and proposed a complexity-invariant distance measure (CID)

\[ CID(X, Y) = \|X - Y\| \frac{\max(CE(X), CE(Y))}{\min(CE(X), CE(Y))} \]

where \( CE(X) = \|X_2^N - X_1^{N-1}\| \) estimates distance between the time series and its lagged counterpart.

**Limitations:** do not take subsequences into account but compare series pairwise.

6. FEATURE CLUSTERING

The approaches include clustering of autocorrelation functions, Fourier and wavelet transformation, dimensionality reduction, and other transformations of raw time series.

Vlachos et al. [30] proposed running the k-means clustering algorithm on the level $i$ of Haar wavelet representation of the data, projecting the final centers obtained for level $i$ from space $2^i$ to space $2^{i+1}$ for level $i + 1$. If any time series were swapped between clusters then repeat previous steps.

Fu et al. [31] proposed time series smoothing and dimensionality reduction prior to clustering with self-organizing maps. Only the Perpetually Important Points that match predefined Querying points are kept. $D(PIP, Q) = \mathbb{E} \left[ (T(PIP) - T(Q))^2 \right]$.

**Limitations:** the approaches are very specific and require data engineering.

6. MODEL-BASED CLUSTERING

The main assumption for these approaches is that there exists an underlying generating process that can be modeled with certain models (e.g. ARMA and its modifications).

Maharaj [32] proposed a clustering approach based on the $\chi^2$ test statistic. They assumed that time series $X$ and $Y$ are generated by an autoregressive process of order $k$, AR($k$) with parameters $\theta^X = \{\theta^X_1, ..., \theta^X_k\}$ and $\theta^Y = \{\theta^Y_1, ..., \theta^Y_k\}$. Setting the null hypothesis $H_0: \theta^X = \theta^Y$ they cluster $X$ and $Y$ together if the $p$-value is greater than the predefined threshold.

**Disadvantage:** The main weakness of the approach is due to the simplicity and linearity of the AR($k$) model. In general it might be hard to select the appropriate model.

7. FORECASTING. LEMPEL-ZIV AND MARKOV CHAINS

LZW algorithm partitions time series $X$ into a collection of subsequences $S^0, S^1, ..., S^M$. $S^k$ starts at time $k$ and represents the shortest previously unobserved subsequence. The prediction is made as

$$P(X_{t+1} = \beta | X^t_1) = \frac{c(S^k \beta | X^t_1)}{c(S^k | X^t_1)}$$

where $c(S^k \beta | X^t_1)$ stands for the number of times $S^k$ is followed with $\beta$.

Markov chain predictor assumes $P(X_{t+1} = \beta | X^t_1) = P(X_{t+k+1} = \beta | X^t_{k+1})$

Denoting repeating subsequences $X^t_{t-k+1} = X^t_{t+k} = \{S_1, ..., S_k\} = S$ we predict

$$P(X_{t+1} = \beta | X^t_1) = \frac{c(S \beta | X^t_1)}{c(S | X^t_1)}$$

**Limitations:** LZW and Markov chain-based predictors treat numeric values as symbols and cannot capture actual magnitudes of the values $X_t$ of time series.
We partition time series $X$ into $k$ overlapping subseries of size $p$ creating feature vectors $x \in \mathbb{R}^{(p-1) \times (N-p+1)}$ and label vectors $y \in \mathbb{R}^{N-p+1}$

$$x = (X_1^{p-1}, X_2^p, ..., X_{N-p+1}^{N-1}), \quad y = (X_p^N)$$

Standard models are fit on $x$ and $y$. Support Vector Machines [33] (SVM) minimize

$$\frac{1}{N-p+1} \sum_{i=1}^{N-p+1} \max \left(0, 1 - y_i (w^T x_i - b)\right) + \lambda \|w\|^2$$

Basic neural networks (NN) fit the model [34]

$$f(x_i) = w_2^T \phi (w_1^T x_i + b_1) + b_2$$

**Limitations:** machine learning approach does not take into account order within each vector $x_i$.

7. FORECASTING. ARIMA MODELS

ARIMA models [34] seem to be the most natural predictors for time series. They take into account autoregression and integrated moving average. ARIMA is defined as

\[(1 - \sum_{i=1}^{p} \phi_i L^i)(1 - L)^d X_t = \delta + (1 + \sum_{i=1}^{q} \theta_i L^i) \epsilon_t\]

where \(\theta, \phi\) are the moving average and autoregression parameters, \(p\) and \(q\) define the lag of the autoregression and moving averages, \(d\) defines integration, \(\epsilon\) stands for normally distributed error, \(L^k X_t = X_{t-k}\), \(\delta\) regulates the drift of the model.

The parameters are often selected with Akaike Information Criterion (AIC) [35]

\[\text{AIC}(p, d, q) = -2 \log(\mathcal{L}) + 2(p + d + k + 1)\]

where \(\mathcal{L}\) is the maximum likelihood estimator, \(k = 1\) if there is a constant term.

**Limitations:** does not capture non-linear relations within time series.

Recurrent Neural Network (RNN) [36] models are among the first to capture temporal dependencies within the subsequences of time series. A fully interconnected model

\[
X_t = \sum_{i=1}^{l} W_i g_i(t)
\]

**Limitations:** the models are too complex and impractical to store long subsequences.

\[
g_i(t) = f \left( \sum_{j=1}^{\text{MAX}(p,q)} \tilde{w}_{ij} X_{t-j} + \sum_{k=1}^{q} \sum_{l=1}^{l} \tilde{w}_{ilk} g_l(t - k) + \theta_i \right)
\]

7. FORECASTING. LSTM

Long Short-Term Memory (LSTM) [37] NN introduced neurons with internal structure. LSTM contain 3 gates: input, output, and forget. 2 input gates

\[ i_1 = \tanh(b_{i1} + X_t W_{i1}^i + h_{t-1} W_{i1}^h), \]
\[ i_2 = \sigma(b_{i2} + X_t W_{i2}^i + h_{t-1} W_{i2}^h) \]

1 forget gate

\[ f = \sigma(b_f + X_t W_{f}^f + h_{t-1} W_{f}^h) \]

and 1 output gate

\[ o = \sigma(b_o + X_t W_{o}^o + h_{t-1} W_{o}^h) \]

Overall, the model is a combination

\[ h_t = \tanh(f + i_1 \otimes i_2) \otimes o \]

SUMMARY

1. Combinatorial estimators of complexity and entropy, as well as predictors including Lempel-Ziv approximator, permutation entropy, Markov chains and Lempel-Ziv predictors treat time series like a sequence of symbols and do not take actual values into account.

2. Methods quantifying predictability rely on entropy or assumed models.

3. Dependency measures efficient at capturing non-linear patterns use transformations that are not guaranteed to have inverses.

4. Majority of the distance measures proposed for time series clustering compare values pairwise and do not take subsequencies into account.

5. Machine learning approach to clustering and time series forecasting does not take order within features into account.
1. Creating new non-linear dependency measures with transformations that are guaranteed to have inverses. (Developing approximations for maximum and monotone correlation).

2. Generalizing combinatorial methods for complexity and entropy estimation so they take magnitudes of time series into account.

3. Combining dependency measures with neural networks. Is there a link and can one improve another?
THANK YOU!

QUESTIONS?