

MolTool package

Andrea Coletta, PhD

25, September 2015

1 TimeSer

1.1 Algorithms

1.1.1 Digitalize

We usually deal with a data series $\vec{x}(t) \in \mathbb{R}^d$ where $d \in \mathbb{N}$ (the number of dimensions of the singular value) and $t \in \mathbb{N}$ (the time-frame at which the value is observed).

In some cases we need to deal with a “digitalized” version of the time-series, where the data are encoded in an integer format. Obviously this encoding is not unique but it depends on a particular “code”. In this case we encode the data using a value-grid defined using the same bins used to calculate the histograms. Let’s assume, for sake of simplicity that the number of bins used to perform the digitalization is the same in each dimension.

$$\begin{aligned}\vec{x}(t) &\in \mathbb{R}^d \text{ (the original data)} \\ d &= \text{number of dimensions} \\ N_b &= \text{number of bins} \\ B &= \{B_0, \dots, B_{N_b}\} \text{ (bins)}\end{aligned}$$

$$\begin{aligned}\text{Digitalize : } \vec{x} \in \mathbb{R}^d &\longmapsto \vec{b} \in \mathbb{N}^d \\ b_i &= \max_{j \in \{0, \dots, N_b\}} (j | B_j \leq x_i) - 1 \\ i &\in \{1, \dots, d\}\end{aligned}$$

1.1.2 Reduce to 1-D

Given a binning grid, the original data series $\vec{x}(t) \in \mathbb{R}^d$ it could be reduced to a 1-dim integer series. The new series is an “encoded” version: the values univocally identify the particular bin of the binning grid, in which the original value falls.

$$\begin{aligned}\vec{x}(t) &\in \mathbb{R}^d \text{ (the original data)} \\ d &= \text{number of dimensions} \\ N_b &= \text{number of bins} \\ B &= \{B_0, \dots, B_{N_b}\} \text{ (bins)}\end{aligned}$$

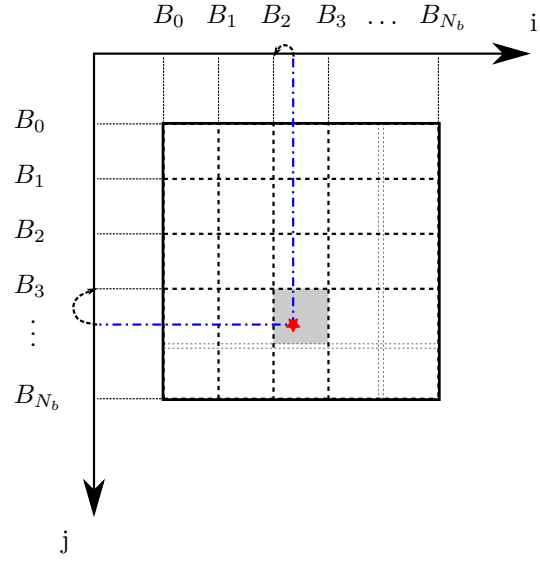


Figure 1: Graphical representation of the “Digitalize” aglorithm.

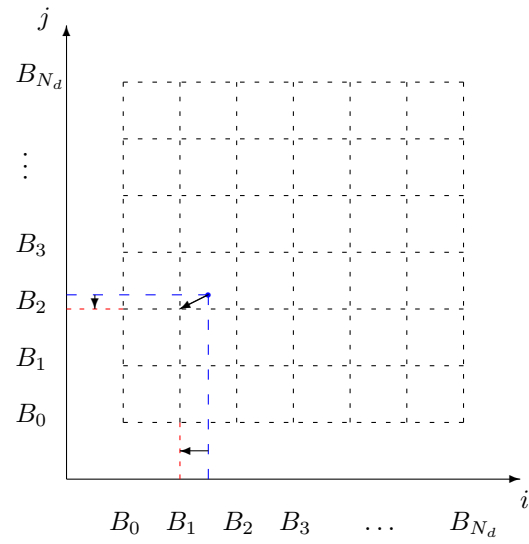


Figure 2: Graphical representation of the “Digitalize” aglorithm.

Also in this case, for sake of simplicity, we use the same binning in each dimension.

$$\begin{aligned} \text{Digitalize} : \vec{x} \in \mathbb{R}^d &\longmapsto b \in \mathbb{N} \\ b &= \sum_{i=0}^{d-1} \underbrace{\left[\max_{j \in \{0, \dots, N_b\}} (j | B_j \leq x_i) - 1 \right]}_{\text{see "Digitalize" sec.1.1.1}} * N_b^i \\ i &\in \{1, \dots, d\} \end{aligned}$$

1.1.3 Trajectory

In order to calculate the entropy rate of a time series, and with it also to estimate the entropy transfer into another time series, starting from the original $\vec{x}(t) \in R^d$ data we need to create an accessory time series in which the singular element represents a time window of a pre-defined number of frames. The original data is previously “encoded” into a “digitized” form (using the “Digitalize” algorithm, sec. 1.1.1) and subsequently a hash number is calculated for each time window biunivocally corresponding to one of the possible value combinations:

$$\begin{aligned} \vec{b}(t) &\in \mathbb{N}^d \text{ (the digitalized data)} \\ \tau &\in \mathbb{N} \text{ (width of the time windows used)} \\ d &= \text{number of dimensions} \\ N_b &= \text{number of bins} \\ B &= \{B_0, \dots, B_{N_b}\} \text{ (bins)} \end{aligned}$$

The bins values correspond also to the only possible value that each component of \vec{b} can assume. Also in this case we assume that the number of bin used in each dimension is the same.

So the function which starting from the original time-series generates a new time-series for which each value correspond to a different subsequence of length τ is:

$$\begin{aligned} \text{Traj} : \vec{b}(t) \in \mathbb{R}^d &\longmapsto w(t) \in \mathbb{N} \\ w(t) &= \sum_{k=0}^{\tau-1} \underbrace{\left[\sum_{i=0}^{d-1} \overbrace{d_i(l+k)}^{\text{"digitalized"}} * N_b^i \right]}_{\text{1-dim reduction}} * N_b^{k*d} \end{aligned}$$