Article

Matrix Profile

Roberto Chiosa1

1 Department of Energy “Galileo Ferraris”, TEBE research group, BAEDA Lab, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Turin, Italy

1. Properties

* ﻿It is exact: the Matrix Profile based methods provide no false positives or false dismissals. It can handle missing data: Even in the presence of missing data, we can provide answers which are guaranteed to have no false negatives.
* It is simple and parameter-free: In contrast, the more general algorithms in this space that typically require building and tuning spatial access methods and/or hash functions.
* It is space efficient: Matrix Profile construction algorithms requires an inconsequential space overhead, just linear in the time series length with a small constant factor, allowing massive datasets to be processed in main memory (for most data mining, disk is death).
* It is incrementally maintainable: Having computed the Matrix Profile for a dataset, we can incrementally update it very efficiently. In many domains this means we can effectively maintain exact joins/motifs/discords on streaming data forever. MP is extremely scalable, for extremely large datasets we can compute the Matrix Profile in an anytime fashion, allowing ultra-fast approximate solutions and real-time data interaction.
* Simplicity and Intuitiveness: Seeing the world through the MP lens often invites/suggests simple and elegant solutions.
* It can be constructed in deterministic time: given only the length of the time series, we can precisely predict in advance how long it will take to compute the Matrix Profile. (this allows resource planning)
* It can leverage hardware: Matrix Profile construction is embarrassingly parallelizable, both on multicore processors, GPUs, distributed systems etc.

1. Definitions

* **Definition 1:** A **time series** T ∈ ℝn is a sequence of real-valued numbers ti ∈ ℝ : T = {t1,t2 … tn} , where n is the length of T
* ﻿**Definition 2:** A **subsequence** Ti,m ∈ ℝm is a continuous subset of values from T of length m starting from position i. Formally defined as ti ∈ ℝ : Ti,m = {ti,ti+1 … ti+m-1}
* **Definition 3:** A **distance profile** D is a vector of the Euclidean distances between a given query and each subsequence in an all subsequence length. In the original matrix profile formulation ﻿is assumed that the distance is measured using the Euclidean distance between the z-normalized subsequences
* **Definition 4:** An **all-subsequences set** A of a time series T is an ordered set of all possible subsequences of T obtained by sliding a window of length m across T : A = {T1,m , T2,m … Tn-m+1,m} where m is a user-defined subsequence length. We use A[i] to denote Ti,m
* **Definition 5:** Given two all-subsequences sets A and B and two subsequences A[i] and B[j], a **1NN-join function** θ1NN(A[i], B[j]) is a Boolean function which returns “true” only if B[j] is the nearest neighbor of A[i] in the set B.
* **Definition 6:** Given all-subsequences sets A and B, a **similarity join set JAB** of A and B is a set containing pairs of each subsequence in A with its nearest neighbor in B : JAB = {〈 A[i], B[j] 〉 | θ1NN(A[i], B[j]). We denote this formally as JAB=A⋈ θ1NN B.
* **Definition 7:** A **matrix profile** (or just profile) PAB is a vector of the Euclidean distances between each pair in JAB . We call this vector the matrix profile because one (inefficient) way to compute it would be to compute the full distance matrix of all the subsequences in one time series with all the subsequence in another time series and extract the smallest value in each row (the smallest non-diagonal value for the self-join case).
* **Definition 9:** The ith element in the matrix profile tells us the Euclidean distance to the nearest neighbor of the subsequence of T, starting at i. However, it does not tell us where that neighbor is located. This information is recorded in matrix profile index. A **matrix profile index IAB** of a similarity join set JAB is a vector of integers where IAB[i] = j if {A[i], B[j]} ∈ JAB

Immagine che contiene freccia

Descrizione generata automaticamente

Politecnico di torino total power, left and right matrix profile

1. Limitations of MP application in energy field

In energy field the time series presents different periodicity (day, week, month, year) resulting different set of window size. The most interesting period is the daily one.

* Problem: the MP does recognize weekend as discords (high values) -> discords will always be on weekend and motif during weekdays.
* Possible solution: introducing expert knowledge, i.e. annotation vector when searching motif/discords

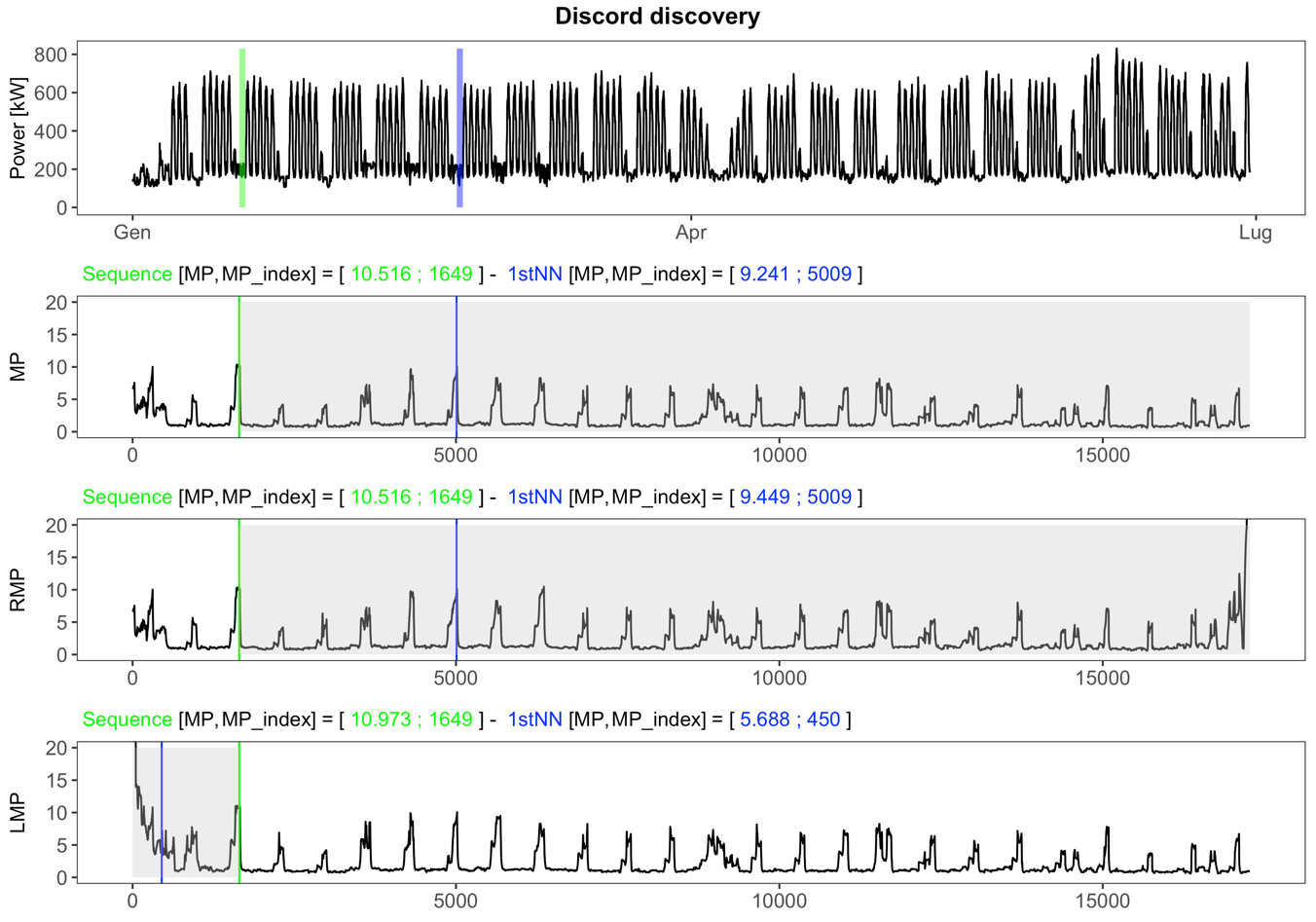
An energy anomaly not only consists into a different load shape but also different magnitude (same profile but higher electrical load).

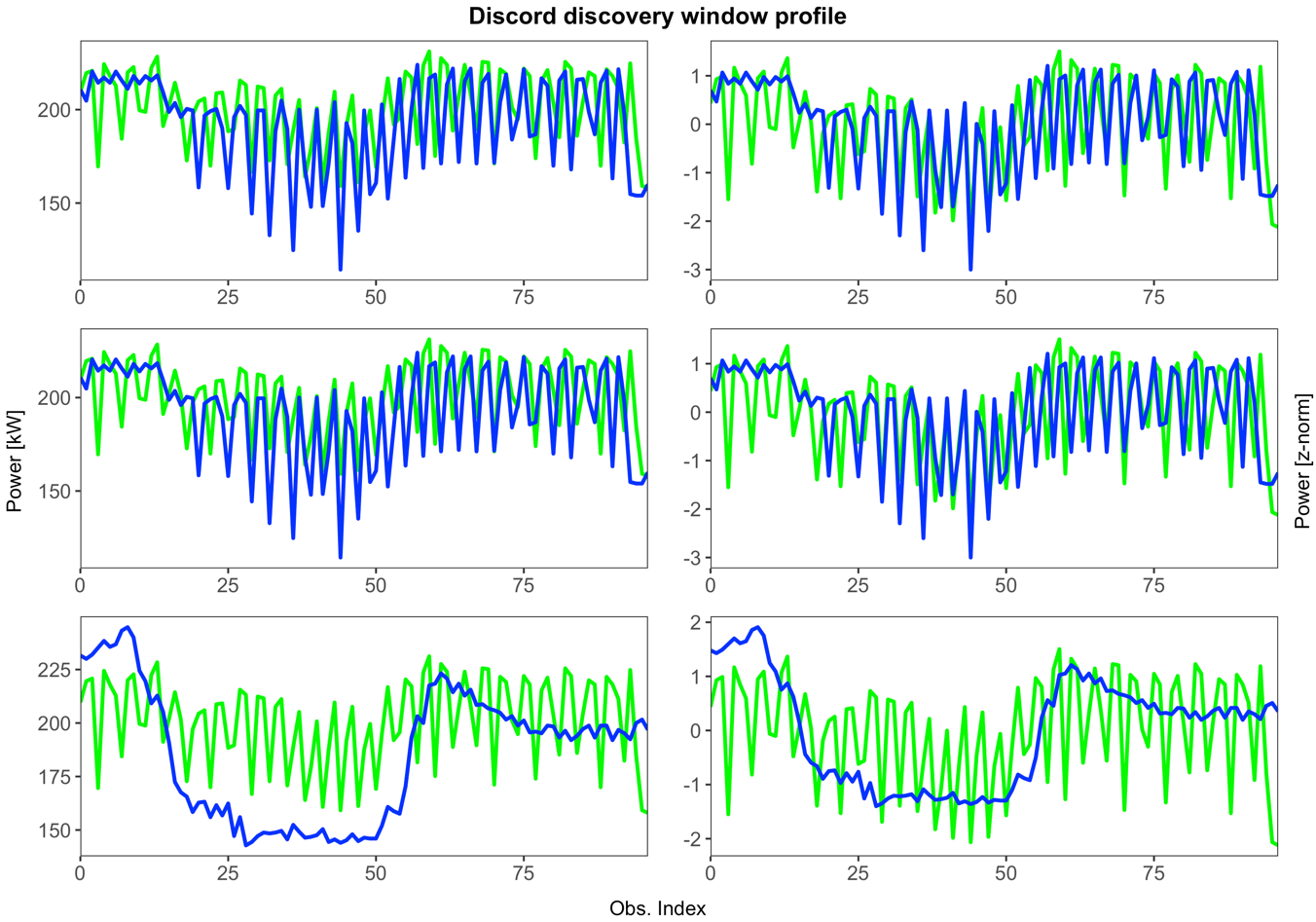
* Problem: The classic MP (with z-score normalization) may fail to identify these kind of anomaly while is effective in identifying anomalous shapes. MP searches NN based on shape similarity.
* Possible solution: avoid z-score normalization of sub-sequences and use pure Euclidean distance

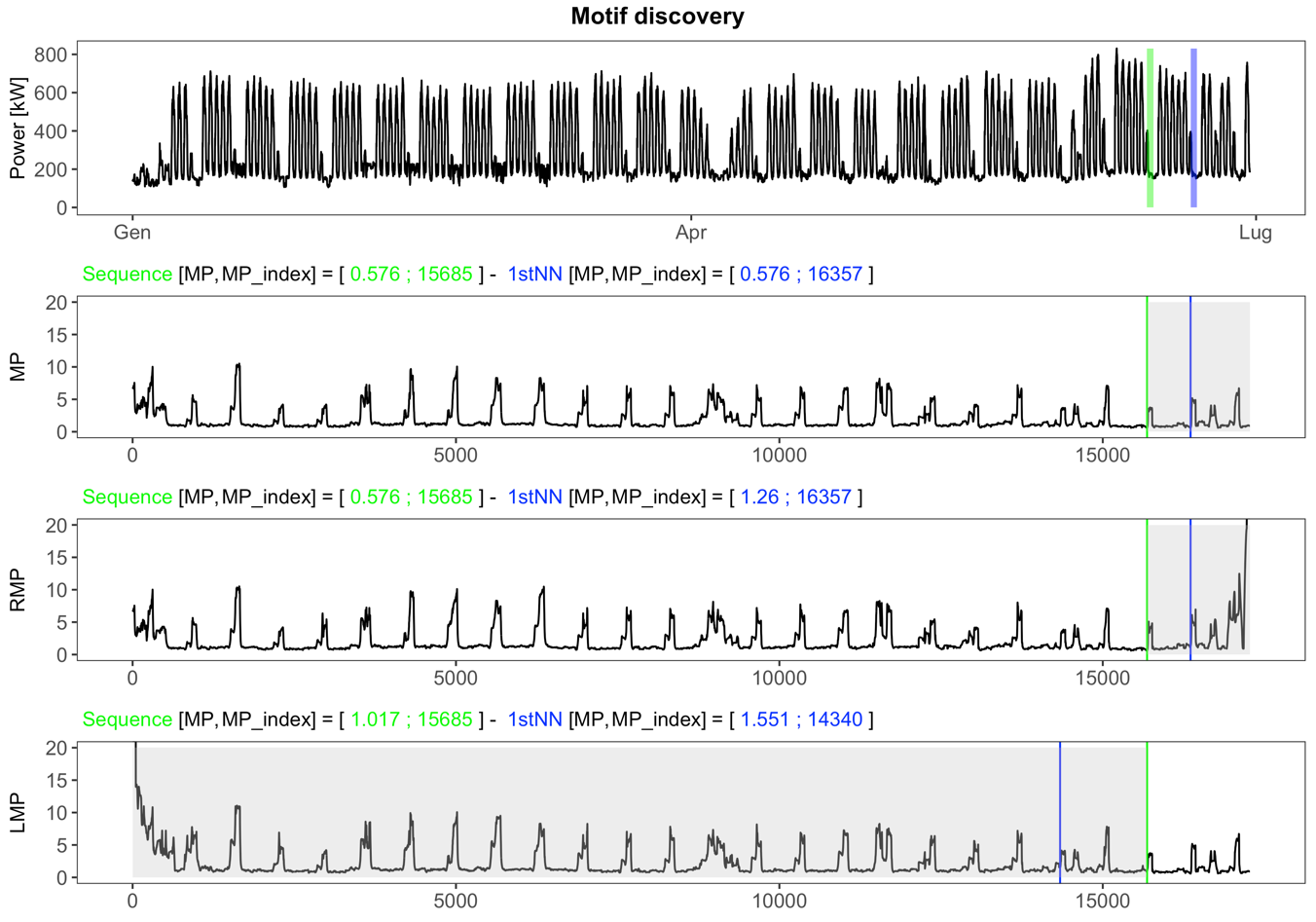
1. Open research path

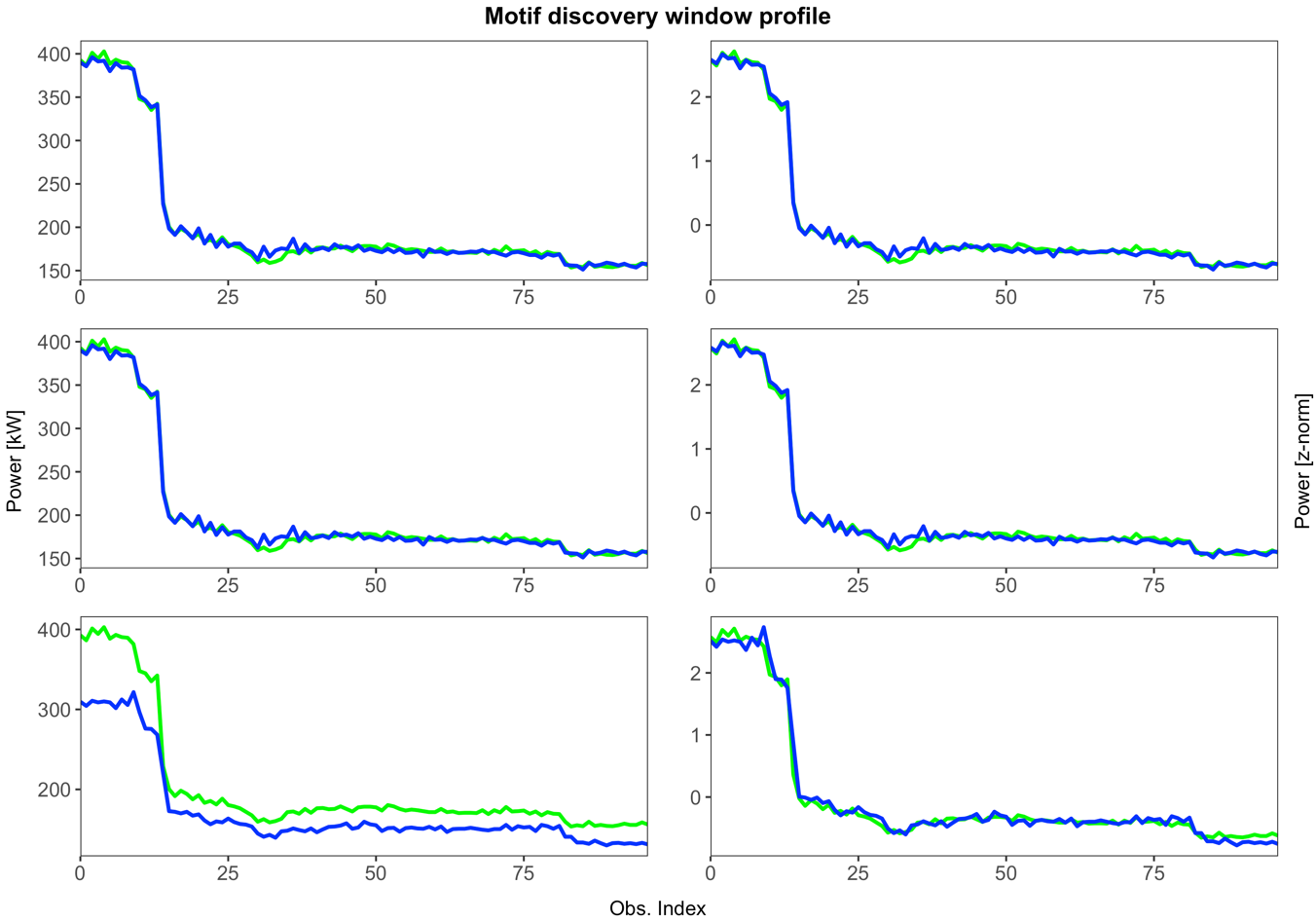
* Comparison with other techniques PAA SAX. MP does not loose information during the process, is it more effective in finding anomalies than dimensionality reduction methods?
* Is there a way to overcame the z-score normalization issues in finding energy anomaly? Is the Euclidean distance more effective and which are the limitations?
* Incorporate domain knowledge in motif/discord/sequence discovery (could be combined with 2)
* To adapt the method in the energy field, makes sense to apply MP on detrended time series?
* Automatic real time/streaming identification of motif discord. Fast and suitable for real time deployment online and continuous learning

1. Motif and discord discovery









1. Annotation Vector

The Annotation Vector (AV) plays the role of manipulating the motif search. ﻿Our main idea is to leverage this MP to discover more meaningful motifs. We achieve this goal by combining the matrix profile with the annotation vector to produce a new matrix profile. We will refer to this as the “Corrected” MP (CMP), as it correctly incorporates the contextual bias for the problem at hand.

Real valued Annotation Vector

The annotation vector AV is a time series consisting of real-valued numbers between [0 - 1]. A low value indicates the subsequence starting at that index is not a desirable motif, and therefore should be biased against. Conversely, higher values mean the subsequence at that location should be favored for the potential motif pool. Note that the AV has the same length as the matrix profile MP.

﻿CMP\_i = MP\_i + (1 - AV\_i) \* max(MP)

Boolean Annotation Vector

Note that here we created a Boolean AV, which strictly prohibits finding motifs during sensor movement. However, we could also have created a real-valued AV, which simply bias the motif search away from regions where movement was noted, in proportion to the magnitude of the motion.

﻿Hard-Limited Artifacts

This is not reflective of medical reality, but is simply a region where the physical process exceeds the 8-bit precision available to record it [21].

We slide a window across the time series to extract subsequences, counting the number of constant values (from being hard-limited above or below) in each sequence. This number over the subsequence length is used as the bias function. A higher value hints that the motifs that include this subsequence may be spurious, as the overflow/underflow regions act like a “don’t-care” in the motif distance.

Stop-word

We assume only that the stop-word motif(s) will be known to the users of our framework, as the stop-words are domain specific.

﻿Simplicity bias

To do this, we employ the complexity estimation proposed in [1], which will be referred as CE. The authors of [1] originally embedded CE in a complexity correction factor for the Euclidean distance, making this distance measure complexity-invariant.

We slide a window across the time series, measuring the complexity of each subsequence and store them in a complexity vector,

Immagine che contiene testo

Descrizione generata automaticamente

Motion Artifact.

Immagine che contiene testo

Descrizione generata automaticamente