Multivariate Motif Discovery with Motiflets

Bachelor Thesis Exposé

Huseyn-zada, Niyaz

Supervisors:
Dr. rer. nat. Patrick Schäfer
Prof. Dr. Matthias Weidlich

September 20, 2023

Contents

1 Introduction **1**

2 Background **2**
2.1 Basic Term Definitions 2
2.2 Distance Measure Definitions 2
2.3 Motif Definitions 3
2.4 Matrix Profile Definitions 3
2.5 Comparison of Matrix Profile and Motiflets 4

3 Related Work **5**

4 Objectives **5**

5 Methods **6**
5.1 Theoretical Foundation (I) 6
5.2 Algorithm Development (II) 6
5.3 Implementation (III) 6
5.4 Evaluation (IV) 7
5.5 Comparison (V) 7
5.6 Documentation (VI) 7

1 Introduction

Time series analysis is a critical component in a multitude of fields, ranging from financial analysis to health monitoring. A key aspect of time series analysis is the detection of motifs, recurring patterns that can provide significant insights into the underlying system. The Motiflets technique, as described in the paper "Motiflets - Simple and Accurate Detection of Motifs in Time Series" [1], has proven effective for motif set detection in univariate time series. This technique defines motifs as the set of exactly k occurrences of a motif of length l, whose maximum pairwise distance is minimal. This approach has shown quantitative and qualitative superiority, leading to clearer and easier to interpret motifs.

However, real-world data often exists in multiple dimensions, and the complexity of analyzing such multivariate time series is significantly higher as they capture multiple co-occurring patterns across various dimensions. The paper "Matrix Profile VI: Meaningful Multidimensional Motif Discovery" [2] presents the mSTAMP approach to apply pair motif discovery algorithms to multivariate time series. It introduces the concept of a Matrix Profile, a meta time series that stores the z-normalized Euclidean distance between each subsequence and its first-nearest neighbor subsequence, and discusses its applications in multivariate motif discovery. Figure 1 provides an illustrative example of a multivariate time series, highlighting a multivariate pair motif.

The aim of this thesis is to extend the Motiflets technique for the detection of motif sets in multivariate time series. By combining the strengths of the Motiflets technique and the multivariate pair motif discovery methods presented in the aforementioned papers, this work seeks to develop a more robust and effective approach for motif set detection in multivariate time series. The potential impact of this research is significant, as improved motif detection techniques can lead to more accurate and insightful analyses in various fields that rely on time series data.

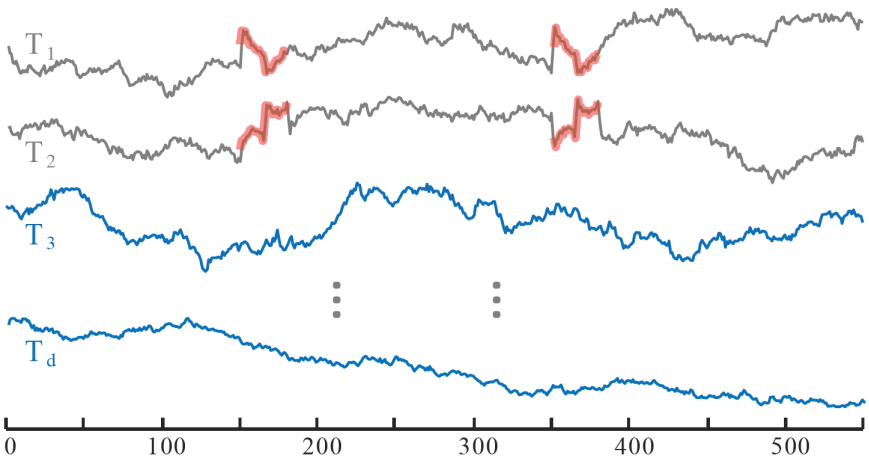

Figure 1: An example of a multivariate time series. Both of the first two dimensions have a pair motif of length 30 embedded at location 150 and 350. All remaining time series (just two are shown above) are simply random walks [2].

2 Background

In the next section, we provide important definitions related to time series analysis and motif detection.

2.1 Basic Term Definitions

It's important to understand some basic terms in time series analysis. First, we'll define what exactly is meant by a univariate time series.

- **Definition 2.1.1 Univariate time series:** A univariate time series $T = (t_1, t_2, ..., t_n)$ of length n is an ordered sequence of n real-values $t_i \in \mathbb{R}$.

  After defining univariate time series, the next step is to consider its segments. We will now explore how a portion or subsequence of a univariate time series can be represented.

- **Definition 2.1.2 Univariate subsequence:** A univariate subsequence $S_{i;l}$ of a univariate time series $T = (t_1, ..., t_n)$, with $1 \leq i \leq n$ and $1 \leq i + l \leq n$, is a univariate time series of length l, consisting of the l contiguous real-values of T starting at offset i: $S_{i;l} = (t_i, t_{i+1}, ..., t_{i+l-1})$.

  Building on our understanding of univariate subsequences and univariate time series, we turn our attention to more complex multivariate time series.

- **Definition 2.1.3 Multivariate time series:** A multivariate time series $\boldsymbol{T} \in \mathbb{R}^{d \times n}$ is a set of co-evolving time series $T^{(i)} \in \mathbb{R}^d : \boldsymbol{T} = [T^{(1)}, T^{(2)}, ..., T^{(d)}]^T$ where d is the dimensionality of $\boldsymbol{T}$ and n is the length of $\boldsymbol{T}$.

  Now that we've established the concept of a multivariate time series, it's essential to understand its components, just as we explored univariate subsequences in univariate time series.

- **Definition 2.1.4 Multivariate subsequence:** A multivariate subsequence $\boldsymbol{S_{i;l}} \in \mathbb{R}^{d \times l}$ of a multivariate time series $\boldsymbol{T}$ is a set of subsequences from $\boldsymbol{T}$ of length l starting from position i. Formally, $\boldsymbol{S_{i;l}} = [S_{i;l}^{(1)}, S_{i;l}^{(2)}, ..., S_{i;l}^{(d)}]^T$.

  Having understood how subsequences are represented in multivariate time series, we further explore scenarios where only specific dimensions are of interest. This leads us to the concept of submultivariate subsequences.

- **Definition 2.1.5 Submultivariate subsequence:** A submultivariate (subdimensional) subsequence $\boldsymbol{S_{i;l}}(X) \in \mathbb{R}^{k \times l}$ is a multivariate subsequence for which only a subset of dimensions is selected, where X is an indicator vector that shows which dimension is included, and k is the number of dimensions included (i.e., $|X| = k$).

2.2 Distance Measure Definitions

After establishing foundational concepts about time series and their subsequences, it becomes important to understand how we measure the similarity or difference between these subsequences. This brings us to the concept of z-normalized Euclidean distance.

- **Definition 2.2.1 z-normalized Euclidean distance (z-ED):** Given two univariate subsequences $S_{a,l} = (s_{a,1}, ..., s_{a,l})$ with mean μ_a and standard-deviation σ_a and $S_{b,l} = (s_{b,1}, ..., s_{b,l})$ with μ_b and σ_b, both of length l, their z-normalized Euclidean distance (*z-ED*) is defined as:

$$z\text{-}ED(S_{a,l}, S_{b,l}) = \sqrt{\sum_{s=1}^{l} (\frac{s_{a,t} - \mu_a}{\sigma_a} - \frac{s_{b,t} - \mu_b}{\sigma_b})^2}.$$

Once we understand how to compute the z-normalized Euclidean distance for univariate subsequences, we can extend this knowledge to multivariate scenarios. This introduces us to the k-dimensional distance function.

- **Definition 2.2.2 k-dimensional distance function [2]:** The k-dimensional distance function or $z\text{-}ED^{(k)}$ computes the distance between two multivariate subsequences by considering only the k out of d dimensions that provide the most significant contribution to the overall distance between the multivariate subsequences. Formally,

$$z\text{-}ED^{(k)}(\boldsymbol{S_{i;l}}, \boldsymbol{S_{j;l}}) := \min_{X \in \mathcal{P}_k(\{1,2,\ldots,d)\}} z\text{-}ED(\boldsymbol{S_{i;l}}(X), \boldsymbol{S_{j;l}}(X)),$$

where $|X| = k$.

However, it's crucial to comprehend the combinatorial implications while thinking about a naive method for choosing dimensions. The number of ways to choose a subset of k dimensions out of d total dimensions is given by the binomial coefficient $\binom{d}{k}$. This represents a combinatorial explosion and underscores the complexity of such naive approaches.

2.3 Motif Definitions

Exploring the methods of measuring distances between subsequences in both univariate and multivariate contexts, our attention now turns to identifying pattern sets within univariate time series. This introduces us to the concept of top k-Motiflets. We begin with the *extent* definition, essential for understanding the top k-Motiflets definition.

- **Definition 2.3.1 Extent [1]:** Consider a univariate time series T and a set S of univariate subsequences of T of length l. The extent of S is the maximal pairwise distance of elements from S:

$$d = extent(S) = \max_{(S_a, S_b) \in S \times S} z\text{-}ED(S_a, S_b).$$

Having defined the concept of *extent*, we can now proceed to define the top k-Motiflets.

- **Definition 2.3.2 Top k-Motiflet [1]:** Given a univariate time series T, cardinality $k \in \mathbb{N}$ and length l, the top k-Motiflet of T is the set S with $|S| = k$ univariate subsequences of T of length l for which the following holds: All elements of S are pairwise d-matching, with $d = extent(S)$, and there exists no set S' with $extent(S') < extent(S)$ also fulfilling these constraints.

2.4 Matrix Profile Definitions

Now that we have addressed motifs, we will define the matrix profile, a tool designed to assist in identifying pair motifs.

- **Definition 2.4.1 Matrix profile [3, 4]:** A matrix profile $P \in \mathbb{R}^{n-l+1}$ of a univariate time series T is a meta time series that stores the z-normalized Euclidean distance between each univariate subsequence and its first-nearest neighbor subsequence, where n is the length of T, and l is the given subsequence length. An example of a matrix profile is shown in Figure 2.

After we defined the matrix profile, it's essential to understand its multivariate counterpart. We will next define the k-dimensional matrix profile.

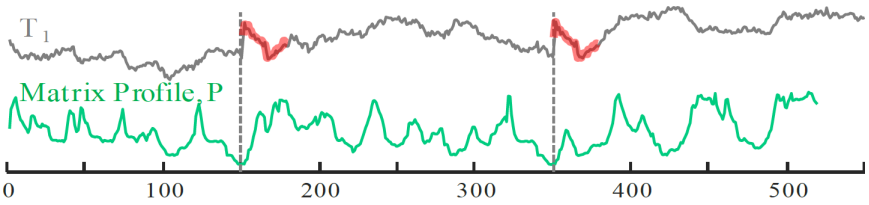

Figure 2: Matrix profile of T_1. The two lowest points on P correspond to the locations of embedded motif pair (red) [2].

- **Definition 2.4.2 k-dimensional matrix profile [2]:** A k-dimensional matrix profile $\boldsymbol{P} \in \mathbb{R}^{n-l+1}$ of a multivariate time series $\boldsymbol{T}$ is a meta time series that stores the z-normalized Euclidean distance between each univariate subsequence and its first-nearest neighbor subsequence (the distance is computed using k-dimensional distance function), where n is the length of $\boldsymbol{T}$, d is the dimensionality of $\boldsymbol{T}$, $k \leq d$ is the given number of dimension, and l is the given subsequence length. An example of a k-dimensional matrix profile for various values of k is illustrated in Figure 3. Formally, the i^{th} element of $\boldsymbol{P}$ stores:

$$z\text{-}ED^{(k)}(\boldsymbol{S_{i;l}}, \boldsymbol{S_{j;l}}) \forall j \in [1, 2, ..., n - l + 1], \text{ where } i \neq j.$$

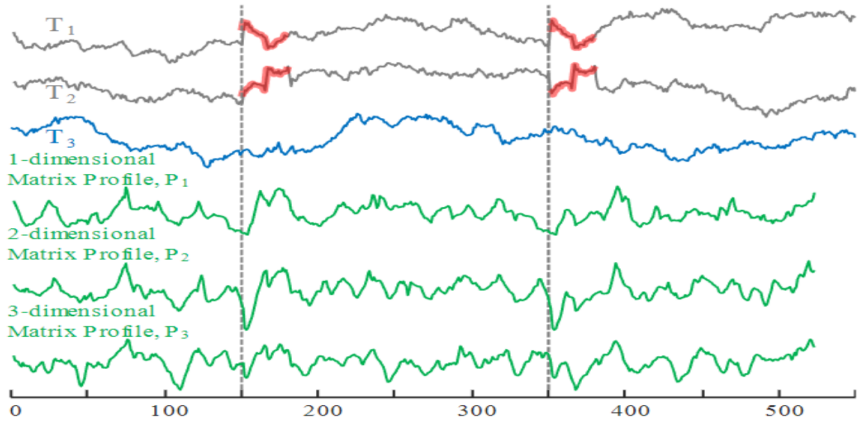

Figure 3: An example of a k-dimensional matrix profile for all possible values of k [2].

2.5 Comparison of Matrix Profile and Motiflets

The Matrix Profile and Motiflets both provide methods for finding motifs in time series, but they do it in different ways. The Matrix Profile provides a meta-time series detailing the z-normalized Euclidean distance between each subsequence and its first-nearest neighboring subsequence. Motiflets, in contrast, concentrate on finding motif sets in univariate time series. Notably, the distance matrix is quickly calculated for both methods using mSTAMP.

Although the Matrix Profile is not directly utilized in this work, it remains a valuable tool for motif discovery in time series, and is worth describing. For the purpose of this thesis, focus is placed on the distance matrix, with the Matrix Profile not being directly applied.

3 Related Work

Time series data analysis has been a key research area for many years because of its importance in fields like finance, health, and engineering. Finding motifs, which are repeated patterns in time series data, is essential as they help us understand the data's underlying patterns.

The Motiflets method, discussed in the paper "Motiflets - Simple and Accurate Detection of Motifs in Time Series" [1], is a new way to find motif sets in one-dimensional time series. This method is simple and accurate, identifying motifs by looking at how often they appear and their distances from each other. The Motiflets method has shown to be better than other methods in many ways, making it a strong choice for finding motif sets.

However, when we look at multivariate time series data, detecting k out of d dimensions get more complicated. The paper "Matrix Profile VI: Meaningful Multidimensional Motif Discovery" [2] addresses this by showing how to find motif pairs in multivariate time series. It's essential to note that this paper focuses only on the discovery of motif pairs. It introduces the Matrix Profile, which measures the distance between parts of the time series and their closest matches. This helps in finding important motif pairs in data with multiple dimensions, giving a full picture of patterns in different areas.

Both papers give important information about finding motif sets or motif pairs in time series data. But there's still a need to expand the Motiflets method for multivariate data. This thesis plans to do that by using parts of both methods, aiming to create a strong method for finding motif sets in multivariate time series.

4 Objectives

The main goal of this thesis is to expand the Motiflets method, which is known to effectively detect motif sets in univariate time series, so it can work with multivariate time series. The detailed goals of this thesis include:

I **Theoretical Foundation:** Develop a theoretical framework for applying the Motiflets technique to multivariate time series. This involves understanding the underlying principles of the Motiflets technique and the challenges of multivariate time series analysis, and devising strategies to integrate the two.

II **Algorithm Development:** Based on the theoretical framework, design an algorithm that implements the extended Motiflets technique. This algorithm should be capable of detecting motifs in multivariate time series effectively and efficiently.

III **Implementation:** Implement the designed algorithm using Python as the chosen programming language. The Python implementation should be robust, efficient, and user-friendly, allowing others to use and further develop it.

IV **Evaluation:** Evaluate the performance of the implemented algorithm on various multivariate time series datasets. The evaluation should assess the algorithm's effectiveness in motif detection, its computational efficiency, and its advantages and disadvantages compared to existing techniques.

V **Comparison:** Compare the extended Motiflets technique with existing techniques for multivariate motif detection, such as the Matrix Profile method. This comparison should clarify the advantages and disadvantages of each method and provide insight into whether or not it is appropriate for use with various kinds of multivariate time series.

VI **Documentation:** Document the entire process, from the development of the theoretical framework to the implementation and evaluation of the algorithm.

By achieving these goals, this thesis aims to advance the area of time series analysis by offering a method for motif set discovery in multivariate time series that is more reliable and efficient. This might improve the usefulness and quality of analysis in many disciplines that use multivariate time series data, including finance, health monitoring, and many others.

5 Methods

The approaches that will be used to accomplish the goals of this thesis are described in this section. The methods are informed by the insights from the papers "Motiflets - Simple and Accurate Detection of Motifs in Time Series" [1] and "Matrix Profile VI: Meaningful Multidimensional Motif Discovery" [2].

5.1 Theoretical Foundation (I)

- **Literature Review:** A comprehensive review of the paper [1] will be conducted to understand the Motiflets technique in depth. Additionally, other relevant literature, including the paper [2], will be reviewed to grasp the challenges and techniques of multivariate time series analysis.

5.2 Algorithm Development (II)

- **Motiflets Extension:** An initial prototype will be developed to extend the Motiflets technique for multivariate time series. It is expected that this modification will not be difficult.

- **Channel Selection:** A more challenging aspect is the implementation of channel selection. When analyzing multivariate time series data, not all dimensions or channels might be equally relevant or useful. In order to minimize noise or redundant information, an effective channel selection technique can help in appropriately evaluating the relevance and significance of each channel. This can lower the cost of calculation and significantly improve the quality of the analysis.

- **Optimization:** Feedback from the prototyping phase will be used to refine and optimize the algorithm, ensuring its effectiveness and efficiency in detecting motifs in multivariate time series.

5.3 Implementation (III)

- **Tool Selection:** Python, the same language in which the original Motiflets is implemented, will be the primary choice for this implementation due to its versatility in data analysis. Relevant libraries, such as NumPy [5] and Numba [6], will be employed to enhance the efficiency of the algorithm. For data visualization, libraries like Matplotlib [7] and Seaborn [8] will be used.

- **Coding:** The extended Motiflets algorithm will be implemented following the specifications determined during the prototyping and optimization phases.

5.4 Evaluation (IV)

- **Dataset Selection:** For evaluation purposes, we will prioritize using well-known multivariate time series datasets such as pop songs, multivariate penguin motion, and in general, multivariate motion capture (MOCAP) data.
- **Performance Metrics:** Metrics like efficiency will be utilized to evaluate the algorithm's performance. Existing methods, such as the Matrix Profile approach, will be used to compare the outcomes.

5.5 Comparison (V)

- **Analysis:** A detailed analysis will be conducted to compare the strengths and weaknesses of the extended Motiflets technique against other methods.

5.6 Documentation (VI)

- **Methodology:** The entire process, from literature review to implementation, will be documented in detail.

List of Figures

1 An example of a multivariate time series. Both of the first two dimensions have a pair motif of length 30 embedded at location 150 and 350. All remaining time series (just two are shown above) are simply random walks [2]. 1
2 Matrix profile of T_1. The two lowest points on P correspond to the locations of embedded motif pair (red) [2]. 4
3 An example of a k-dimensional matrix profile for all possible values of k [2]. . . . 4

References

[1] Patrick Schäfer and Ulf Leser. Motiflets - Simple and Accurate Detection of Motifs in Time Series. *Proceedings of the VLDB Endowment*, pages 725–736, 2022. URL `https://www.vldb.org/pvldb/vol16/p725-schafer.pdf`.

[2] Chin-Chia Michael Yeh, Nickolas Kavantzas, and Eamonn Keogh. Matrix Profile VI: Meaningful Multidimensional Motif Discovery. *IEEE International Conference on Data Mining (ICDM)*, 2017. URL `https://www.cs.ucr.edu/~eamonn/Motif_Discovery_ICDM.pdf`.

[3] Chin-Chia Michael Yeh, H. Van Herle, and Eamonn Keogh. Matrix Profile III: The Matrix Profile Allows Visualization of Salient Subsequences in Massive Time Series. In *IEEE International Conference on Data Mining (ICDM)*, 2016.

[4] Yan Zhu, Zachary Zimmerman, Nima S. Senobari, Chin-Chia Michael Yeh, Gareth Funning, Abdullah Mueen, Philip Brisk, and Eamonn Keogh. Matrix Profile II: Exploiting a Novel Algorithm and GPUs to break the one Hundred Million Barrier for Time Series Motifs and Joins. In *IEEE International Conference on Data Mining (ICDM)*, 2016.

[5] Oliphant, Travis. NumPy, 2023. URL `https://numpy.org/`.

[6] Oliphant, Travis. Numba, 2023. URL `https://numba.pydata.org/`.

[7] Hunter, John D. Matplotlib, 2023. URL https://matplotlib.org/.

[8] Michael Waskom. Seaborn, 2023. URL https://seaborn.pydata.org/.