The Swiss Army Knife of Time Series Data Mining: Ten Useful Things you can do with the Matrix Profile and Ten Lines of Code

Yan Zhu

Diego Silva, Shaghayegh Gharghabi, Eamonn Keogh (author order to be decided based on contributions, but Yan first and Eamonn last)

University of California, Riverside

mueen@unm.edu, eamonn@cs.ucr.edu

Abstract— The recently introduced data structure, the Matrix Profile, annotates a time series by recording the location *of* and distance *to* the nearest neighbor to every subsequence. This information trivially gives the answer to both *time series motifs* and *time series discords*, perhaps the two most frequently used primitives in time series data mining. One attractive feature of the Matrix Profile is that is completely divorces the high-level details of the analytics performed, from the computational "heavy lifting". The Matrix Profile can be computed using the appropriate computational paradigm, CPU, GPU, FPGA, distributed computing, anytime computation, incremental computation, etc., but this can all be hidden from the analyst. Expanding on this philosophy, in this work we ask the following question. If we assume that we get the Matrix Profile for free, what analytics can we do, writing at most ten lines of code? As we will show, the answer is surprising large and diverse. We can both reproduce the results of many much more complicated algorithms, and find novel regularities in time series.

***Keywords**: Time series, Joins, Motif Discovery, Anomaly Detection*

1 Introduction

The (ek will do)

The original Matrix Profile paper concludes with the sentence, "*There are many avenues for future work, and we suspect that the research community will find many uses for, and properties of, the matrix profile that did not occur to us*." []. With the Matrix Profile having spent a year it the public eye, we are now ready to consider uses for the Matrix Profile.

2 General Related Work and Background

To do…

3 Ten Useful Things you can do with the Matrix Profile and Ten Lines of Code

1.1 Discovering Motifs Under Uniform Scaling

The utility for motif discovery under uniform-scaling invariance was first considered in [yy]. We revisit the motivation with a simple and visually compelling example. We took two exemplars from the same class from the MALLET dataset [4], and imbedded them into a random walk dataset. As Figure 1.*top* shows, even without the color-coded clue brushed onto the data by the Matrix Profile discovery tool [], the repeated pattern is visually obvious.

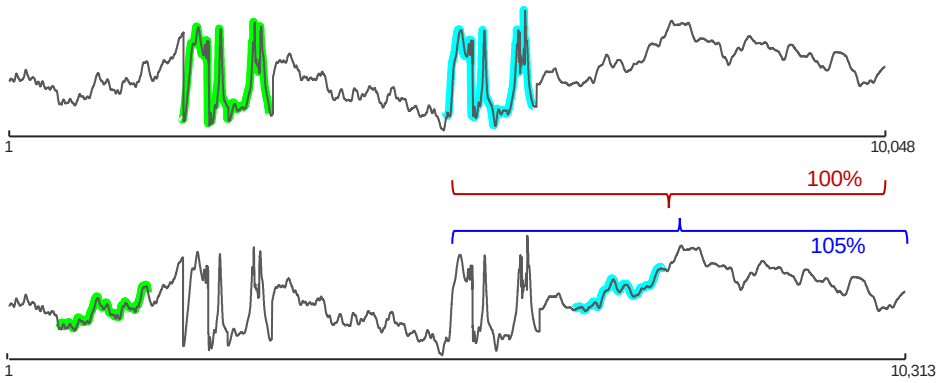

Figure 1: *top*) A random walk dataset with two exemplars from the MALLAT dataset imbedded at locations 2001 and 5025. The color highlighting indicates the top-1 motif, which unsurprisingly are exactly the imbedded patterns. *bottom*) The same dataset, but with the second half linearly stretched by 5%. This causes the top-motif to change to snippets of random walk.

We then took the second half of the time series and linearly stretched it by 5%. By any standard such a change is a trivial difference and essentially visually imperceptible. Nevertheless, as Figure 1.*bottom* shows, the pair of imbedded patterns are no longer the top-1 motif, an unexpected and disquieting result. Before we show how to address this

within this paper's "the Matrix Profile plus ten-lines-of-code framework", we note the following facts that mitigate the issue.

- For the rescaled version, the pair of imbedded patterns was the *second*-best motif, and only *just* nudged out by the spurious random walk pair.
- If, instead of searching with a motif length of 1,024, the original length of the imbedded pattern, we had searched for a shorter length, say 500, then the best motif would have been a subsequence of the imbedded pattern. The user could then have examined the shorter motif, and realized it could be extended significantly while maintaining its similarity.
- We deliberately chose this dataset, from the eighty-five in the UCR archive, knowing it would be very sensitive to changes in linear scaling. This is because *complex* time series (see [zz]) with very sharp rises and falls are particularly sensitive to having features out of phase. For most datasets, motif discovery is much more robust to small amounts of uniform scaling.

Despite all these mitigating facts, Figure 1.*bottom* clearly shows that there may be some situations in which there is a need to find motifs with invariance to uniform scaling. To the best of our knowledge, there is only one research effort that has addressed this, however this method is *approximate*, requires many parameters to be set, and only able

to support a limited range of scaling [yy]. In contrast, we can easily solve this problem *exactly*, under our simple assumptions.

For the moment assume that we know the scaling factor we want to be invariant to happens to be 1.64. We can take the dataset T and copy a stretched version of it into T2, simply by using:

```
T2 = T(1: 100/164: end); % Unofficial way to resample
```

If we now call:

```
[JMP, JMPindex] = computeMatrixProfileJoin(T,T2,500);
```

Now the resulting Matrix Profile will discover the motifs with the appropriate uniform scaling invariance. In fact, we did exactly this on a 6,106,456 length trace of household electrical demand to discover the motif shown in Figure 2.

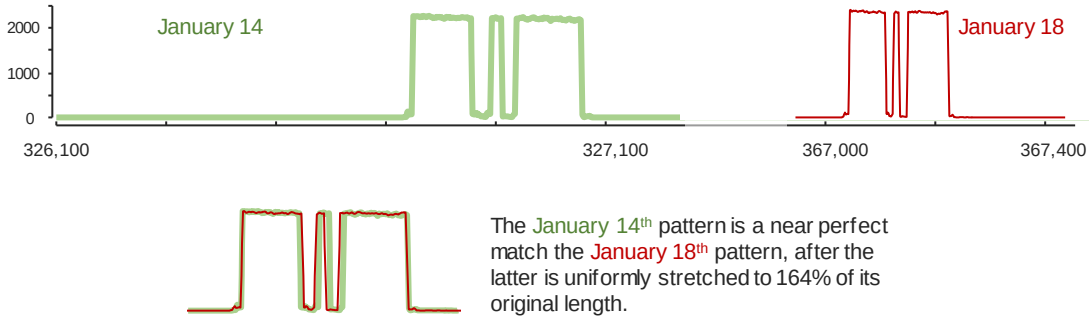

Figure 2: *top*) Two non-contagious snippets from the ElectriSense dataset [aa]. While semantically similar, they have a very large Euclidian distance. *bottom*) After stretching the January 18th pattern by 164%, the two patterns are almost identical.

The motif pattern appears to be the three elements of a dishwasher cycle (clean, rinse, dry), which can take different amounts of time due to the use of the optional *half-load* feature [bb]. In this case, we knew from some first principle physics how to set the scaling factor, but that may not always be the case. However, given our assumptions, we can simply iterate over all possible scaling factors in a given range. For example, to

discover motifs that are similar after scaling one pattern by 150 to 180%, we can use the following code snippet.

```
for scale_factor = 150 : 180
  T2 = T(1: 100/scale_factor: end);
  [JMP, JMPindex] = computeMatrixProfileJoin(T,T2,500);
  <trivial code to record best motifs omitted>
end
```

This example perfectly elucidates the philosophy driving this paper. For many time series data mining tasks, we may not need to spend significant human time designing, implementing and tuning new algorithms. The Matrix Profile and ten lines of code may really be sufficient.

1.2 Discovering Time Series Semordnilaps

Consider the sentence fragment we discovered in Wikipedia, "*..the longest-lived Tasmanian devil recorded was Coolah*.."[xx]. This snippet contains a Semiordnilap pair [cc], the mirrored words "lived" and "devil". Semiordnilaps are easy to find in arbitrary text strings, and indeed have an important role in molecular biology. For example, many restriction enzymes recognize specific palindromic sequences and cut them. As a concrete example, the restriction enzyme EcoR1 recognizes the following palindromic pair, "GAATTC" and "CTTAAG" [dd].

Because the original definition of time series motifs was directly inspired by the analogy to DNA, it is natural to ask if there is a natural time series analogy to semiordnilaps, and if so, can they be efficiently discovered? From the previous example, the reader will readily see that this trivial, we can simply use:

```
T2 = fliplr(T);  % returns T reversed

[JMP, JMPindex] = computeMatrixProfileJoin(T,T2,m);
```

The only question remaining is are there natural domains that contain time series semiordnilaps? The answer is affirmative.

To demonstrate the utility Semordnilap discovery, we consider Joseph Haydn's Symphony No. 47 in G major, written in 1772. In particular, we examined a performance by the Tafelmusik Orchestra, directed by Bruno Weil in 1993 [xxx]. The performance is twenty-one minutes and two seconds long. As shown in Figure 3.*top*,

we converted it to Mel-frequency cepstral coefficients (MFCC) using windows with 0.5 second and 50% of overlap (standard music processing settings). We set *m* to 150, or 37.5 seconds.

At time 14 minutes and 53 seconds there is a Semordnilap of a passage we encountered at 14 minutes and 16 seconds.

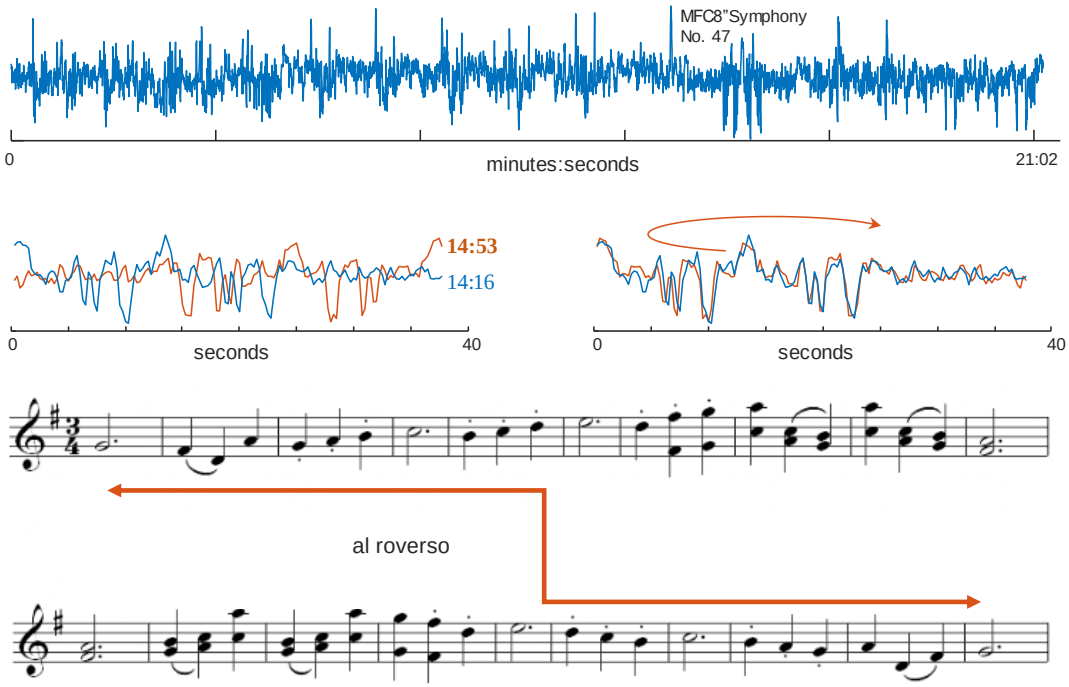

Figure 3: *top*) Haydn's Symphony No. 47 converted to MFCC. *center.left*) Two snippets found by Semordnilap discovery appear unrelated until we flip one backwards in time (*center.right*). *bottom*) The sheet music for the relevant section explains this unexpected discovery.

Figure 3.*bottom* explains the presence of such a perfectly conserved Semordnilap. As noted in [yyy], "*The most extraordinary of all canonic movements from this time is of course from Symphony No. 47. Here Haydn writes out only one reprise of a two-reprise form, and the performer must play the music 'backward' the second time around*".

While this example is clearly contrived, there may be Semordnilaps waiting to be discovered in dance, travel trajectories, industrial processes, and a host of domains that have yet to occur to us.

1.3 Discovering Time Series Reverse Complements

Our success in finding Semiordnilaps immediately suggests another specialized type pattern we could search for. Are there examples of patterns which repeat, but in which one pattern is the inverse of the other? That is to say, unlike Semiordnilaps which are

"flipped" in the *time* axis, are there patterns which are flipped upside-down in the *value* axis? We call such patterns Time Series Reverse Complements (TSRCs).

For example, ENSO (El Nino Southern Oscillation) is a phenomenon that is characterized by intermittent negative correlations between the surface temperatures observed near Australia and Pacific Ocean [ee]. However, there are much more quotidian examples. Consider the two-minute snippet of time series shown in Figure 4. It shows the y-axis from a hip-worn accelerometer from the USC Human Motion Database []. As shown in Figure 4.*bottom.left*, the best motif of length twenty seconds is not well conserved, and almost looks like two random subsequences. This is unsurprising, apart from dance or athletic performances, we would not expect human behavior to faithfully repeat over such an extended time scale. However, we also searched for the best TSRC pattern of the same length, and as shown in Figure 4.*bottom.center* and Figure 4.*bottom.right* it is stunningly well conserved.

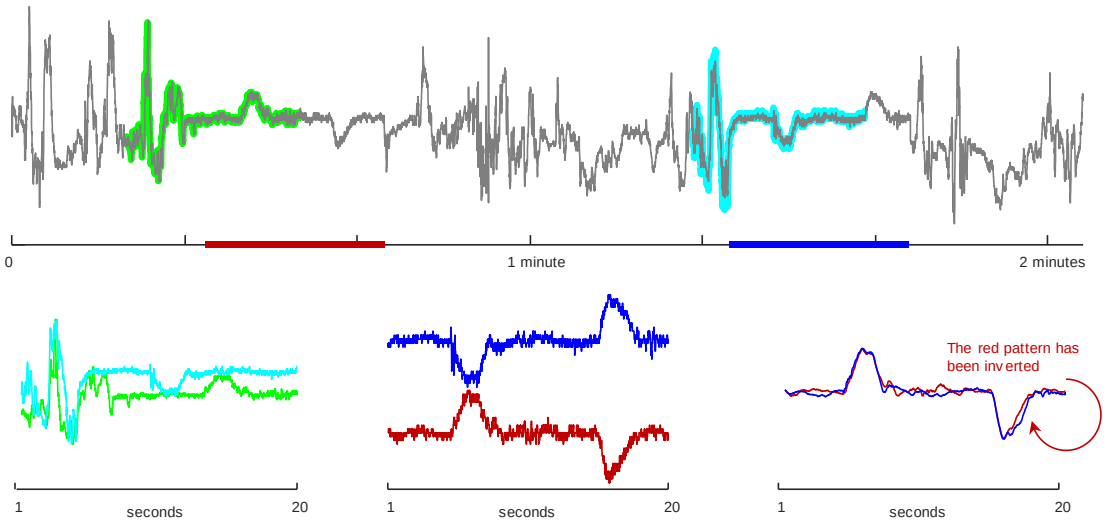

Figure 4: *top*) Approximately two minutes from a dataset from a hip-worn accelerometer of quotidian activity. *bottom.left*) The best motif of length twenty seconds is not well conserved, however, if we generalize the search to consider TSRC motifs (*bottom.center*) we find a highly conserved pattern. To better see how well conserved it is, in (*bottom.right*) we show the patterns with one element inverted, and both patterns smoothed. However, we note that we discovered this pattern in the original noisy space.

What is the mechanism that produced this pattern? At about twenty-two seconds into the recording, the user stepped into an elevator. The first bump is the "jolt" of the elevator ascending, followed by the "dip-and-recover" as the elevator decelerated the

desired floor. After about one minute, the user took a return trip, descending the same number of floors.

The reader will readily appreciate that discovering TSRCs with the matrix profile is trivial, we simply used:

```
T2 = T*-1;  % returns T flipped upside down

[JMP, JMPindex] = computeMatrixProfileJoin(T,T2,m);
```

Note that in this case, the discovered TSRC also happens to be a Semiordnilap. However, this need not be the case in general.

us.

1.4 <more examples>

Placeholder

4 Conclusion

We

Acknowledgements

We gratefully acknowledge funding from…

References

[1] Bayardo, R. J., Ma, Y., and Srikant, R. *Scaling Up All Pairs Similarity Search*. WWW 2007, pp 131-140.

[2] Begum, N., and Keogh, E. *Rare Time Series Motif Discovery from Unbounded Streams*. PVLDB 8(2): 149-160, 2014.

[3] Chandola, V., Cheboli, D., and Kumar, V. *Detecting Anomalies in a Time Series Database*. UMN TR09-004.

[4] Chen, Y. et al. *The UCR Time Series Classification Archive.* http://www.cs.ucr.edu/~eamonn/time_series_data/

[5] Convolution - Wikipedia, the free encyclopedia. https://en.wikipedia.org/wiki/Convolution. Accessed: 2016-01-19.

[6] Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., and Keogh, E. J. *Querying and Mining of Time Series Data: Experimental Comparison of Representations and Distance Measures*. PVLDB 1(2): 1542-1552. 2008.

[7] Y. Li, L. H. U, M. L. Yiu, Z. Gong. Quick-motif: An efficient and scalable framework for exact motif discovery. ICDE 2015: 579-590

[8] Mueen, A., Hamooni, H., and Estrada, T. *Time Series Join on Subsequence Correlation*. IEEE ICDM 2014, pp. 450-459.

[9] Mueen, A., Keogh, E., Zhu, Q., Cash, S. and Westover, B. *Exact Discovery of Time Series Motif*. SDM 2009.

[10] Ueno, K., Xi, X., Keogh, E. J., and Lee, D.-J. *Anytime Classification Using the Nearest Neighbor Algorithm with Applications to Stream Mining*. ICDM 2006.

[11] Ye, L., and Keogh, E. *Time Series Shapelets: A New Primitive for Data Mining*. ACM SIGKDD, 2009. pp 947-56.

[12] Yeh, C.-C. M., Herle, H. V., and Keogh, E. *Matrix Profile III: The Matrix Profile Allows Visualization of Salient Subsequences in Massive Time Series*. To be appeared in IEEE ICDM 2016.

[13] Yeh, C.-C. M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H. A., Silva, D. F., Mueen, A., and Keogh, E. *Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets*. To be appeared in IEEE ICDM 2016.

[14] Zhu, Y., Zimmerman, Z., Senobari, N. S., Yeh, C.-C. M., Funning, G., Mueen, A., Brisk, P., and Keogh, E. *Matrix Profile II: Exploiting a Novel Algorithm and GPUs to Break the One Hundred Million Barrier for Time Series Motifs and Joins*. ICDM 2016.

[15] Zilberstein, S. and Russell, S. *Approximate Reasoning Using Anytime Algorithms*. In Imprecise and Approximate Computation, Kluwer Academic Publishers, 1995.

[xx] Wikipedia entry for Tasmanian devil. Retrieved June 12th 2017. https://en.wikipedia.org/wiki/Tasmanian_devil
[yy] Dragomir Yankov, Eamonn J. Keogh, Jose Medina, Bill Yuan-chi Chiu, Victor B. Zordan: Detecting time series motifs under uniform scaling. KDD 2007: 844-853
[zz] Gustavo E. A. P. A. Batista, Eamonn J. Keogh, Oben Moses Tataw, Vinícius M. A. de Souza:
CID: an efficient complexity-invariant distance for time series. Data Min. Knowl. Discov. 28(3): 634-669 (2014)
[aa] Sidhant Gupta, Matthew S. Reynolds, and Shwetak N. Patel, ElectriSense: single-point sensing using EMI for electrical event detection and classification in the home. In Proceedings of the 12th ACM international conference on Ubiquit

[bb] LG Dishwasher Owners Manual (URL: retrieved on June 21, 2017)
http://www.lg.com/us/support/products/documents/Owners%20Manual.pdf
[cc] Puder, James (2000) "Seventeen Synonyms of Semordnilap," Word Ways: Vol. 33 : Iss. 1 , Article 9.
[dd] Kurpiewski, M. R.; Engler, L. E.; Wozniak, L. A.; Kobylanska, A.; Koziolkiewicz, M.; Stec, W. J.; Jen-Jacobson, L.Mechanism of coupling between DNA recognition specificity and catalysis in EcoRI endonuclease Structure 2004, 12, 1775– 1788
[ee] Kao, Hsun-Ying; Jin-Yi Yu (2009). "Contrasting Eastern-Pacific and Central-Pacific Types of ENSO". J. Climate. 22 (3): 615–632.

[xxx] Music performance of Joseph Haydn's Symphony No. 47 in G major, by the Tafelmusik Orchestra. www.youtube.com/watch?v=yeB_Ohpsm64 Retrieved July 4th 2017.

[yyy] Mark Evan Bonds, “Haydn's 'Cours complet de la composition' and the Sturm und Drang” in Haydn Studies, ed. W. Dean Sutcliffe. Cambridge: Cambridge University Press (1998)