

Multivariate Symbolic Aggregate Approximation for ECG Analysis

Moritz M. Konarski
Supervised by Prof. Taalaibek M. Imanaliev

Applied Mathematics and Informatics Program,
American University of Central Asia

May 3, 2021
Bishkek, Kyrgyz Republic



Outline

1 Introduction

2 SAX

3 MSAX

4 HOT SAX

5 Preliminary Results

6 Outlook

Test

Some Testing

Example

Here we see an example of some bullshit

ECG Basics

- record electrical impulses of the heart
- modern ECGs have 12 leads, online datasets contain 2 or more [1]
- ECGs are multivariate data (multiple data points at each sample point) [2, 3]

$$\mathbf{x}^i[t] = (x_1^i[t], \dots, x_n^i[t]) \quad (1)$$

- ECGs are very common diagnostic tools

ECG waveform

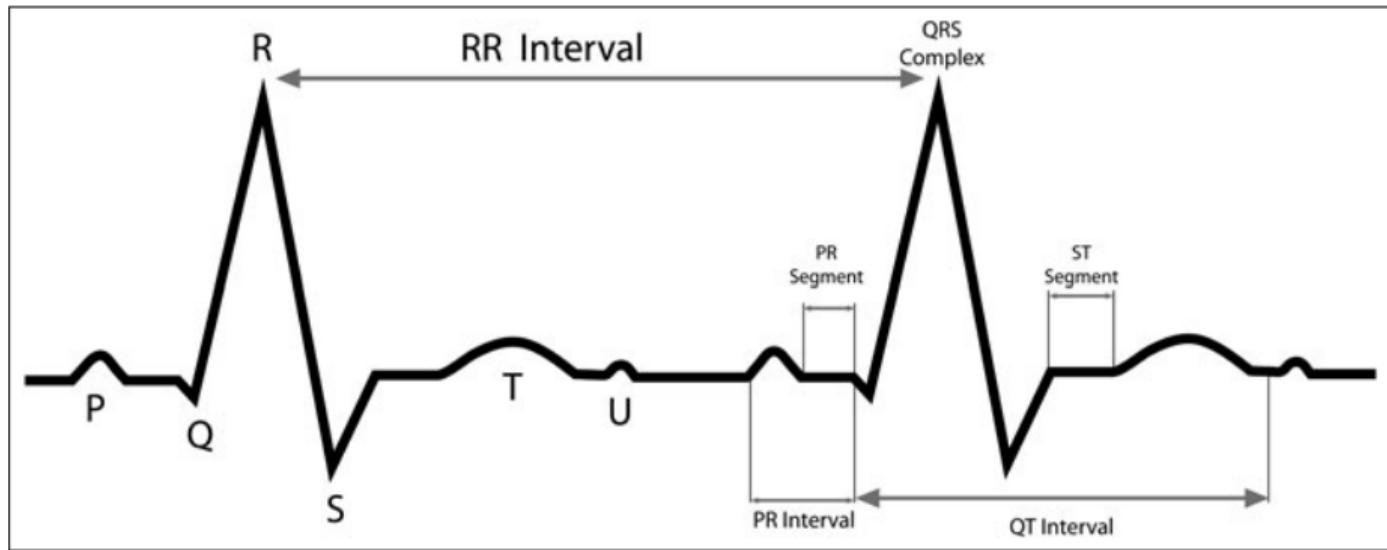


Figure 1: A standard ECG waveform with annotations[4]

Automated ECG Analysis

- ECGs represent large amounts of data, thorough analysis is required
- 5 stages: (1) signal acquisition, filtering; (2) data transformation, processing; (3) waveform recognition; (4) feature extraction; (5) classification[5]
- some methods include FFT, DWT, ANN, kNN, filters
- balance between accuracy and complexity needed

SAX and MSAX

- Symbolic Aggregate Approximation (SAX) creates a simplified, symbolic representation [6]
 - is guaranteed to behave like the original data
 - works on univariate time series, has been used on ECGs[7]
 - Multivariate SAX (MSAX)[3] expands SAX to multivariate time series
- using MSAX on ECGs should increase the accuracy of discord detection compared to SAX

Z-Normalization

- assumption: data is approximately normally distributed
- to analyze time series, they are first normalized so that $\mu = 0$ and $\sigma = 1$

$$x^i[t] = \frac{X^i[t] - \mu}{\sigma} \quad (2)$$

- enables comparison between different time series

PAA

- piecewise aggregate approximation (PAA) reduces dimensionality (through averaging of segments)
- simplifies the time series
- results in $\bar{C} = \bar{c}_1, \dots, \bar{c}_w$
- getting element i of \bar{C} (time series x has length n)

$$\bar{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} x_j$$

PAA Graph

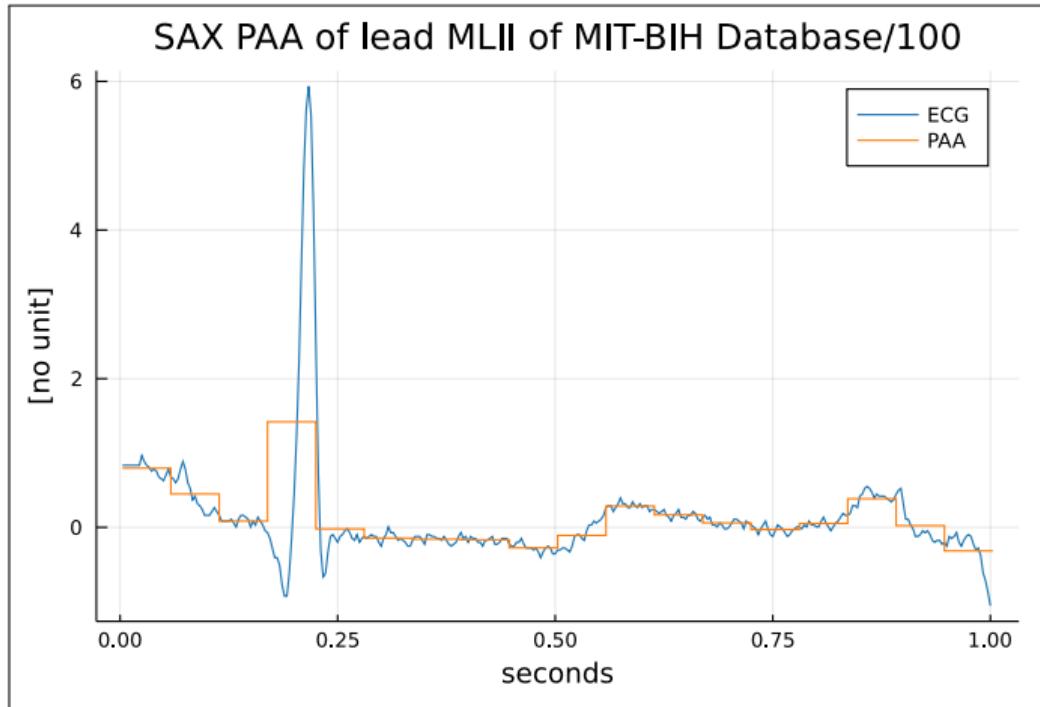


Figure 2: ECG with PAA (MITBIH/100, $w = 18$, $n = 360$)

Discretization

- assign letters to PAA segments
- breakpoints are created that divide a Gaussian curve into equal parts
- number of breakpoints dependent on size of alphabet
- all PAA below lowest breakpoint are a , the ones above it b ...

Discretization Graph

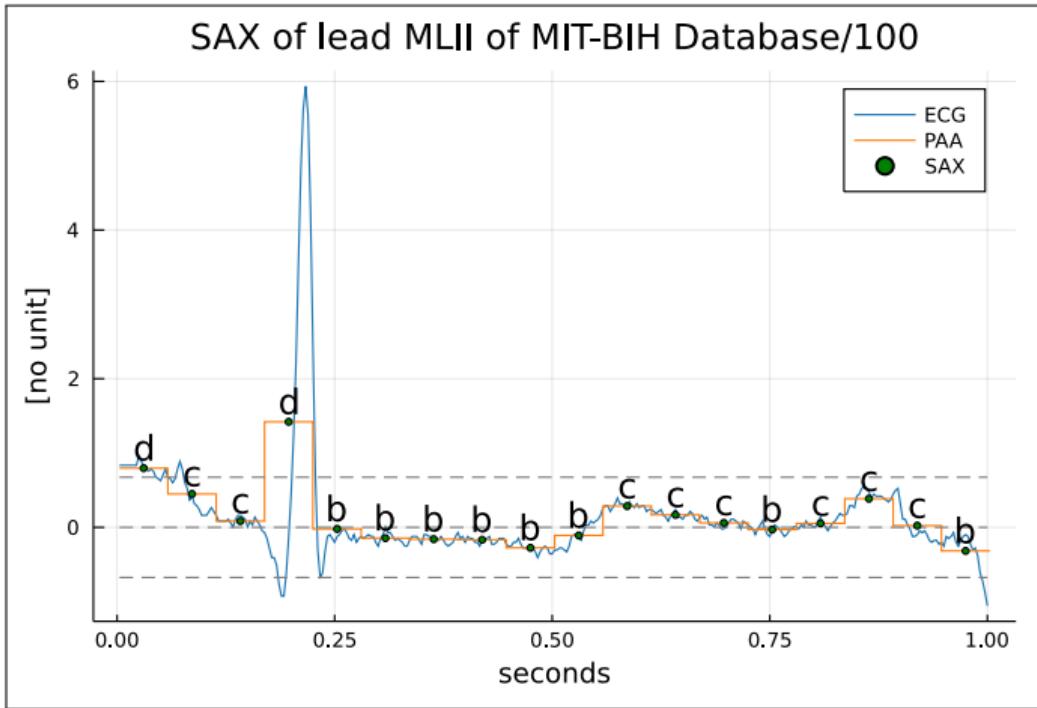


Figure 3: SAX (MITBIH/100, $w = 18$, $n = 360$, alphabet size 3)

Distance Measure

- SAX lower bounds the Euclidean distance, i.e. SAX distances correspond to Euclidean distances
- Euclidean distance between 2 time series Q, C

$$D(Q, C) \equiv \sqrt{\sum_{i=1}^n (q_i - c_1)^2}$$

- SAX distance

$$MINDIST(\hat{Q}, \hat{C}) \equiv \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^w (dist(\hat{q}_i, \hat{c}_i))^2}$$

- $dist(\hat{q}_i, \hat{c}_i)$ is the difference between the breakpoints of \hat{q}_i, \hat{c}_i

Z-Normalization

- perform multivariate normalization
- mean vector μ as vector of the means for each time series
- covariance matrix Σ for variances and covariances between the different time series

$$\mathbf{x}[t] = \Sigma^{-1/2}(\mathbf{X}[t] - \mu)$$

- this uses mean and covariance structure of the multivariate data

PAA and Discretization

- PAA is used here like in SAX, each time series is handled separately
- the discretization process works the same way too
- each time series component is discretized separately
- to differentiate them, one alphabet can for example be uppercase

Discretization Graph

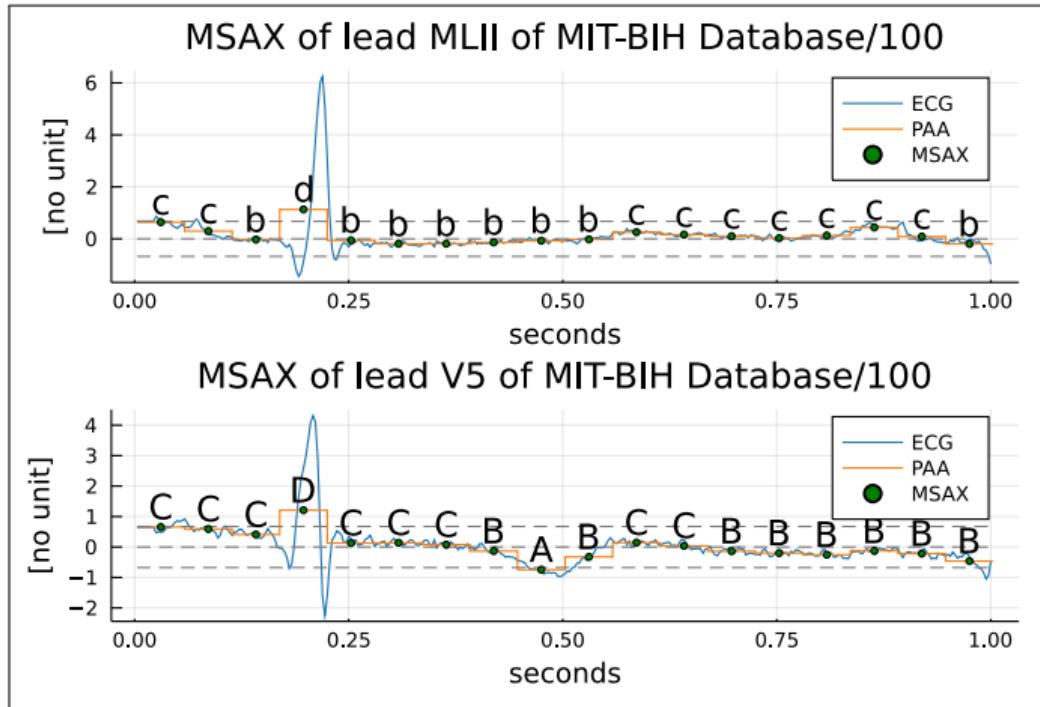


Figure 4: MSAX (MITBIH/100, $w = 18$, $n = 360$, alphabet size 3)

Distance Measure

- this distance measure is based on *MINDIST*
- it is also lower bounding the Euclidean distance
- it adds an extra step of adding the *dist* values for the time series components

$$MINDIST_MSAX(Q, C) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=0}^w \left(\sum_{i=0}^n dist(q[i], c[i])^2 \right)}$$

HOTSAX

- time series discords are sections of a time series that are most different from all other segments of the time series (e.g. diseases in an ECG)
 - can be found by comparing all segments to all other segments[8]
 - Heuristically Ordered Time series using Symbolic Aggregate Approximation is better
 - discords are generally rare, start with the rarest segment
 - similar segments are likely to have similar distances, consider them together
- applying HOTSAX with MSAX to ECGs should discover more discords than HOTSAX with SAX

Preliminary Results

- used the MIT-BIH ECG database
- database has 48 recordings, every heart beat has been annotated by experts
- use HOT SAX with SAX, MSAX to find discords
- use the annotations to check if the discovered discord is a normal heart beat or not
- in ECG 108, HOT SAX with SAX found 16 discords, HOT SAX with MSAX found 21

Outlook

- perform more sophisticated analysis of results
- apply the process to all ECGs in the MIT-BIH database
- experiment with different parameters for SAX/MSAX
- use the method on a different ECG database

References I

- [1] G. B. Moody and R. G. Mark, *MIT-BIH Arrhythmia Database*, 1992. doi: 10.13026/C2F305.
- [2] *Multivariate Data - an overview* | ScienceDirect Topics, [Online]. Available: <https://www.sciencedirect.com/topics/computer-science/multivariate-data> (visited on 03/30/2021).
- [3] M. Anacleto, S. Vinga, and A. M. Carvalho, “MSAX: Multivariate Symbolic Aggregate Approximation for Time Series Classification,” en, in *Computational Intelligence Methods for Bioinformatics and Biostatistics*, P. Cazzaniga, D. Besozzi, I. Merelli, and L. Manzoni, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2020, pp. 90–97, ISBN: 978-3-030-63061-4. doi: 10.1007/978-3-030-63061-4_9.
- [4] J. Wasilewski and L. Poloński, “An Introduction to ECG Interpretation,” en, in *ECG Signal Processing, Classification and Interpretation: A Comprehensive Framework of Computational Intelligence*, A. Gacek and W. Pedrycz, Eds., London: Springer, 2012, pp. 1–20, ISBN: 978-0-85729-868-3. doi: 10.1007/978-0-85729-868-3_1.

References II

- [5] Kligfield Paul *et al.*, “Recommendations for the Standardization and Interpretation of the Electrocardiogram,” *Circulation*, vol. 115, no. 10, pp. 1306–1324, Mar. 2007. doi: 10.1161/CIRCULATIONAHA.106.180200.
- [6] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, “A symbolic representation of time series, with implications for streaming algorithms,” en, in *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery - DMKD '03*, San Diego, California: ACM Press, 2003, pp. 2–11. doi: 10.1145/882082.882086.
- [7] C. Zhang *et al.*, “Anomaly detection in ECG based on trend symbolic aggregate approximation,” en, *Mathematical Biosciences and Engineering*, vol. 16, no. 4, pp. 2154–2167, 2019, ISSN: 1547-1063. doi: 10.3934/mbe.2019105.
- [8] E. Keogh, J. Lin, and A. Fu, “HOT SAX: Efficiently Finding the Most Unusual Time Series Subsequence,” en, in *Fifth IEEE International Conference on Data Mining (ICDM'05)*, Houston, TX, USA: IEEE, 2005, pp. 226–233, ISBN: 978-0-7695-2278-4. doi: 10.1109/ICDM.2005.79.