Automated Screening of Dyslexia via Dynamical Recurrence Analysis of Wearable Sensor Data

Michaela Areti Zervou[1,2], George Tzagkarakis[1], and Panagiotis Tsakalides[1,2]
[1]*Institute of Computer Science, Foundation for Research and Technology-Hellas, Heraklion, Greece*
[2]*Department of Computer Science, University of Crete, Heraklion, Greece*
E-mails: zervou@ics.forth.gr, gtzag@ics.forth.gr, tsakalid@ics.forth.gr

***Abstract*—Dyslexia is a neurodevelopmental learning disorder that affects the acceleration and precision of word recognition, therefore obstructing the reading fluency, as well as text comprehension. Although it is not an oculomotor disease, readers with dyslexia have shown different eye movements than typically developing subjects during text reading. The majority of existing screening techniques for dyslexia's detection employ features associated with the aberrant visual scanning of reading text seen in dyslexia, whilst ignoring completely the behavior of the underlying data generating dynamical system. To address this problem, this work proposes a novel self-tuned architecture for feature extraction by modeling directly the inherent dynamics of wearable sensor data in higher-dimensional phase spaces via multidimensional recurrence quantification analysis (RQA) based on state matrices. Experimental evaluation on real data demonstrates the improved recognition accuracy of our method when compared against its state-of-the-art vector-based RQA counterparts.**

***Index Terms*—Dyslexia screening, multidimensional recurrence quantification analysis, non-linear data analysis, wearable sensors**

I. Introduction

Dyslexia is a neurodevelopmental learning disability that adversely affects between 5-10% of the population [1]. Specifically, it affects the way information is processed, stored and retrieved, with problems of memory, speed of processing, time perception, organization and sequencing [2]. Early-age identification and diagnosis of dyslexia is imperative in order to provide the necessary assistance to dyslexic candidates, since, as individuals grow older, compensatory mechanisms develop that help alleviate the symptoms of dyslexia [3]. However, the learning gap that has been developed is followed by poor school performance, causing psychological and emotional distress, low self-esteem, lack of motivation and depression [4], [5].

Although dyslexia is not a primary oculomotor disease, eye movements differ during reading between typical and dyslexic readers [6], [7]. Typically, in readers with dyslexia, fixation duration and number of fixations increase, average saccade length gets shorter and the number of regressions (i.e., short backward eye movements targeting text that has already been read) increases [8], [9]. The observed differences can be attributed to abnormal linguistic or cognitive processing [10].

The majority of existing screening techniques for dyslexia's detection employ features associated with the aberrant visual scanning of reading text seen in dyslexia [1], [11]. In [12], the one-dimensional counterpart of RQA [13] is applied on dyslexia's data for investigating dyslexic and non-dyslexic word-naming performance in beginning readers. Although such methods can lead to high-precision results in the one-dimensional case for relatively smooth data, they lack the capability of concurrently processing multiple dimensions. To overcome these limitations, this work proposes an alternative approach for the accurate detection of dyslexia, which exploits the temporal variability of the underlying dynamical system that generates the data. To this end, a generalization of the multidimensional recurrence quantification analysis (mdRQA) framework [14] is proposed to perform a sophisticated non-linear analysis of multiple sensor streams by exploiting both the intra- and inter-sensor correlations. Apart from the capability of an mdRQA-based approach to treat non-stationary and short data series, furthermore, it comprises of a set of appropriate quantitative measures for the quantification of recurrent, typically small-scale, structures, and the detection of critical transitions in the systems dynamics (e.g. deterministic, stochastic, random). To the best of our knowledge, there is no prior work in the literature that employs mdRQA to detect dyslexia from multidimensional data.

The contributions of this paper are the following:

(i) the underlying multidimensional data generating processes are modeled concurrently and directly in a higher-dimensional phase space, identifying more accurately the time-evolving dynamics of sensor streams;
(ii) our proposed generalized multidimensional RQA (GmdRQA) method exploits the correlations not only within a stream but also between pairs of streams;
(iii) an efficient feature extraction scheme is designed for the discovery of information-rich patterns that best capture the underlying data dynamics;
(iv) a totally self-tuned architecture is designed for unsupervised dyslexia's detection.

The rest of the paper is organized as follows: In Section II, the dataset employed by our study is overviewed. Section III analyzes our proposed generalized multidimensional RQA

This work is funded by the Hellenic Foundation for Research and Innovation (HFRI) and the General Secretariat for Research and Technology (GSRT) under grant agreement No. 2285 (neuronXnet), and by the Interreg V-A Greece-Cyprus 2014-2020 programme, co-financed by the European Union (ERDF) and National Funds of Greece and Cyprus, under the project SmartWater2020.

2471-7819/19/$31.00 ©2019 IEEE
DOI 10.1109/BIBE.2019.00144

framework, based on state matrices, for feature extraction. Section IV evaluates the performance of our method and compares its accuracy with its vector-based mdRQA counterpart. Finally, Section V summarizes the main outcomes of this work and gives directions for future extensions.

II. Dataset description

The dataset provided by [1] consists of a sample of 97 (76 males and 21 females) high-risk subjects with early identified word decoding difficulties and a control group of 88 (69 males and 19 females) low-risk subjects. These subjects were selected from a larger population of 2165 school children attending second grade (age 8-9). Eye movement recordings are made while the subjects are reading a short natural passage of text adapted to their age. A goggle-based infrared corneal reflection system, namely, the Ober-2TM (Formerly Permobil Meditech, Inc., Woburn, MA), is used to track eye position over time, by sampling the horizontal and vertical position of both eyes at a frequency of 100 Hz. All subjects read one and the same text presented on a single page of white paper with high contrast. The text is distributed over 8 lines and consists of 10 sentences with an average length of 4.6 words.

III. Proposed Architecture

The vector-based version of mdRQA introduced by [14] extracts the underlying dynamics of an ensemble of recorded data streams by mapping the time series in a higher-dimensional phase space of trajectories. More specifically, given a multidimensional time series of length N we reconstruct the corresponding phase space representation as follows,

$$\begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_{N_s} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_{1,1} & \mathbf{x}_{2,1} & \cdots & \mathbf{x}_{D,1} \\ \mathbf{x}_{1,2} & \mathbf{x}_{2,2} & \cdots & \mathbf{x}_{D,2} \\ \vdots & \vdots & \cdots & \vdots \\ \mathbf{x}_{1,N_s} & \mathbf{x}_{2,N_s} & \cdots & \mathbf{x}_{D,N_s} \end{pmatrix} , \quad (1)$$

where D is the number of dimensions of the streams' ensemble, $x_{i,j} = (r_j, r_{j+\tau}, \ldots, r_{j+(m-1)\tau})$, $i = 1, \ldots, D$, $j = 1, \ldots, N_s$, with m being the embedding dimension, τ the delay and $N_s = N - (m-1)\tau$ the number of states.

Recurrence plots (RPs) [13] have been proposed as an advanced graphical technique of visual non-linear data analysis, which reveals all the times of recurrences, that is, when the phase space trajectory of the dynamical system visits roughly the same area in the phase space as shown in Fig. 1. Accordingly, the multidimensional recurrence plot (mdRP) is defined by,

$$\mathbf{mdR}_{i,j} = \Theta\left(\varepsilon - ||\mathbf{v}_i - \mathbf{v}_j||_p\right), \quad i, j = 1, \ldots, N_s \ , \quad (2)$$

where $\mathbf{v}_i$, $\mathbf{v}_j$ denote the state vectors, ε is a threshold, $||\cdot||_p$ denotes a general ℓ_p norm, d is a distance metric and $\Theta(\cdot)$ is the Heaviside step function, whose discrete form is defined by,

$$\Theta(n) = \begin{cases} 1, & \text{if } n \geq 0 \\ 0, & \text{if } n < 0 \end{cases} , \ n \in \mathbb{R} \ . \quad (3)$$

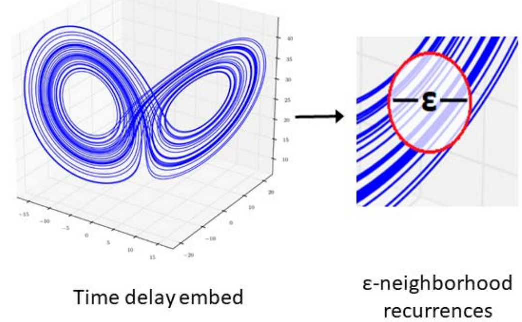

Fig. 1. The time series is time delay embedded into a reconstructed phase space. Then, around each point in the embedded phase space, a recurrence neighbourhood of radius ε is created. All recurrences into this neighbourhood are tracked.

The disadvantage of the conventional mdRQA is that it does not capture the correlations between pairs of distinct streams. To address this limitation, our proposed GmdRQA framework relies on state matrices instead of state vectors, to represent the time-delay embedding of a streams' ensemble. State matrices are considered more appropriate for describing multidimensional signals from a mathematical perspective, enabling them to model the correlations not only within a signal but also between different signals. Specifically, we define a state matrix $\mathbf{X}_i$ as follows,

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{x}_{1,i} & \mathbf{x}_{2,i} & \cdots & \mathbf{x}_{k,i} \\ \mathbf{x}_{k+1,i} & \mathbf{x}_{k+2,i} & \cdots & \mathbf{x}_{2k,i} \\ \mathbf{x}_{2k+1,i} & \mathbf{x}_{2k+2,i} & \cdots & \mathbf{x}_{3k,i} \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{x}_{(l-1)k+1,i} & \mathbf{x}_{(l-1)k+2,i} & \cdots & \mathbf{x}_{lk,i} \end{pmatrix} , \quad (4)$$

where $i = 1, \ldots, N_s$, $k = \lfloor\sqrt{D}\rfloor$ and $l = \lfloor D/k \rfloor$.

In our implementation, the optimal time delay τ is estimated as the first minimum of the average mutual information (AMI) function averaged over all the dimensions in the data [15]. Concerning the embedding dimension m, a minimal sufficient value is estimated using the method of false nearest neighbours (FNN) [16]. Finally, following the empirical rule proposed in [17], the threshold ε is set equal to the 15th percentile of the distribution of all the pairwise distances between the state matrices.

Subsequently, our proposed generalized multidimensional recurrence plot (GmdRP) is defined by

$$\mathbf{GmdR}_{i,j} = \Theta\left(\varepsilon - d(\mathcal{M}(\mathbf{X}_i), \mathcal{M}(\mathbf{X}_j))\right) \ , \quad (5)$$

where ε is a threshold, $\mathcal{M}$ is an operator, d refers to a proper distance metric and $\Theta(\cdot)$ is the Heaviside step function whose discrete form is defined by (3).

As in the one-dimensional case, a major advantage of multidimensional RPs is that they can also be applied to rather short and even non-stationary data. In general, RPs are consisted of isolated points, diagonal as well as vertical lines that form several structures. Therefore, it is often difficult and subjective to analyze. Along these lines, the visual interpretation of RPs, is enhanced by means of several numerical measures for the quantification of the structure and complexity. The following ten measures are utilized to form our feature matrix (ref. [18] for the mathematical definitions):

- **Recurrence rate:** Measures the density of points in the RP or in other words, the probability that a similar state recurs to its neighbourhood in phase space.
- **Determinism:** The ratio of the number of recurrence points forming diagonal structures to the total number of recurrence points is regarded as determinism or predictability of the system. Determinism is close to unity in a periodic system and close to zero in systems with no time-dependence.
- **Average diagonal length:** This average length is actually the mean time that we can predict the next recurrence of states from the state we observe now. Intuitively, a diagonal line of length l means that trajectories are co-evolving during l samples but they correspond to different times of the system evolution. These lines indicate how different trajectories diverge during the evolution of the system and as time passes by.
- **Length of longest diagonal/vertical line:** Refers to the maximum length of the diagonal/vertical lines in the recurrence plot that represent the maximum time that the system evolves or remains in a certain state respectively.
- **Entropy of diagonal/vertical length:** Indicates the complexity of the recurrence plot in respect of the diagonal/vertical lines. The entropy of vertical lines reflects the distribution of time-periods for which the system abides in laminar phases. Signals with no time dependence present diagonal entropy≈0, i.e., the diagonal lines distribution is fully concentrated on very short lines (e.g., single dots).
- **Laminarity:** Provides information about the occurrence of the laminar states in the system. However, it does not describe the length of the laminar states. The value of laminarity decreases if an increased number of single recurrence points are present in the recurrence plot than the vertical structures.
- **Trapping time:** The average length of vertical lines is called trapping time and it is related with laminarity time. This value contains information about the frequency and the length of laminar states.

Finally, a linear-kernel support vector machine (SVM) is applied on the feature matrix for discriminating between the two classes, namely, dyslexia and control. Fig. 2 shows the overall architecture of our proposed GmdRQA-based dyslexia's detection system.

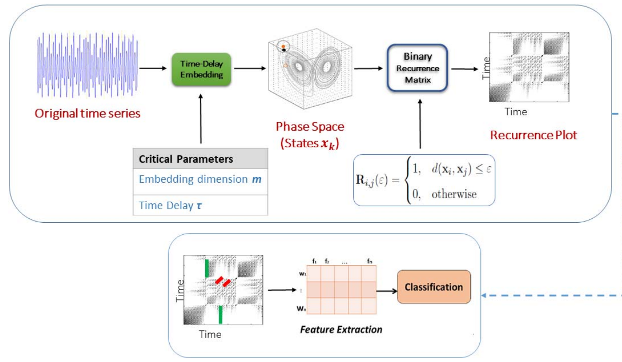

Fig. 2. Proposed generalized GmdRQA-based feature extraction and dyslexia's detection architecture.

TABLE I
CLASSIFICATION ACCURACY AND STANDARD DEVIATION AVERAGED OVER 100 REPETITIONS.

	Mean Classification Accuracy±std (%)
State-of-the-art Vector-based mdRQA	92.30±3.14
Proposed Euclidean Distance Metric of State Matrix Eigenvalues	**92.77±3.32**
Proposed Correlation Matrix Distance Metric	92.39±3.45

IV. Performance Evaluation

A. Distance Metrics and Operators

1) Vector-based mdRQA: The choice of the ℓ_p norm depends on the nature of the data. The most commonly used norms include the (i) Euclidean, (ii) maximum and (iii) minimum norm. Our extensive evaluation on real data showed that the Euclidean norm performs best for the specific type of dyslexia's data.

2) Proposed matrix-based GmdRQA: Given the two-dimensional nature of our state matrices, appropriate distance metrics d must be defined for our GmdRQA method. Specifically, the following distance metrics and matrix operators are utilized and tested for the construction of GmdRPs:

- **Euclidean norm of state matrices eigenvalues:** Setting the operator $\mathcal{M}$ to be the calculation of the eigenvalues vector of a state matrix using the singular value decomposition (SVD) [19], (5) takes the following form,

$$\mathbf{GmdR}_{i,j} = \Theta\left(\varepsilon - \|\mathbf{x}_{eig}^{i} - \mathbf{x}_{eig}^{j}\|_2\right) , \quad (6)$$

  where $\mathbf{x}_{eig}^{i}$ and $\mathbf{x}_{eig}^{j}$ are the eigenvalues vectors of the state matrices $\mathbf{X}_i$ and $\mathbf{X}_j$, respectively, $i, j = 1, \ldots, N_s$.
- **Correlation matrix distance:** In order to measure the change of spatial second-order statistics, the correlation matrix distance (CMD) [20] between correlation matrices is employed, which is defined by

$$d(\mathbf{C}_i, \mathbf{C}_j) = 1 - \frac{tr\{\mathbf{C}_i\mathbf{C}_j\}}{||\mathbf{C}_i||_F||\mathbf{C}_j||_F} \in [0, 1] \quad (7)$$

  where $\mathbf{C}_i$ and $\mathbf{C}_j$ are the correlation matrices of the state matrices $\mathbf{X}_i$ and $\mathbf{X}_j$, respectively, $tr\{\cdot\}$ is the trace operator and $||\cdot||_F$ denotes the Frobenius norm. The CMD becomes zero if the correlation matrices are equal up to a scaling factor and one if they differ to a maximum extent. The more the signal spaces of $\mathbf{C}_i$ and $\mathbf{C}_j$ overlap, the higher becomes the trace of the product and therefore

TABLE II

PRECISION, RECALL AND STANDARD DEVIATION FOR BOTH LOW RISK AND HIGH RISK CLASS AVERAGED OVER 100 REPETITIONS.

State-of-the-art Vector-based mdRQA		
	Low Risk Class (LR)	**High Risk Class (HR)**
Precision	95.80±4.20	89.68±4.54
Recall	87.49±5.21	96.60±3.40

Euclidean distance norm of state matrix eigenvalues		
	Low Risk Class (LR)	**High Risk Class (HR)**
Precision	97.02±2.98	89.61±4.40
Recall	87.23±5.15	97.66±2.34

Correlation matrix disatance		
	Low Risk Class (LR)	**High Risk Class (HR)**
Precision	97.10±2.70	88.99±4.81
Recall	86.39±5.68	97.71±2.18

the CMD decreases. This property of CMD makes it a useful measure to evaluate whether the spatial structure of a signal ensemble, hence, the signals' statistics have changed to a significant amount. Subsequently, by setting $\mathcal{M}(\mathbf{X}_i) = \mathbf{C}_i$, the associated GmdRP is defined by,

$$\mathbf{GmdR}_{i,j} = \Theta\left(\varepsilon - d(\mathbf{C}_i, \mathbf{C}_j)\right) \ , \tag{8}$$

where d is the correlation matrix distance.

B. Classification

The recorded data and metadata of each participant are concatenated and then divided randomly into training and testing subsets containing 70% and 30% of the data, respectively. A non-linear classifier, namely, a linear-kernel SVM, is applied on the generated feature matrix in order to detect dyslexia, which is formulated as a classification problem. The classification process is repeated 100 times and the average performance is reported. The choice of this classifier is motivated by its fast execution, as well as by its high accuracy, especially in the case of a large number of available features. We emphasize, though, that the classification step is decoupled from the feature extraction step, thus the overall performance of the architecture can be further improved by employing a better classifier.

C. Evaluation Metrics

The performance of our generalized matrix-based GmdRQA architecture is compared against its vector-based mdRQA counterpart, in terms of classification accuracy and robustness to noise. Specifically, saccadic eye movement, which is a fast random convulsing movement even when the eye fixates on one point, can be modeled as white noise due to its randomness. Moreover, there is no preference to the direction of the eye movement, therefore, we may claim that the governed eye movement model is approximated by an additive Gaussian noise process. Along these lines, we evaluate the robustness of our proposed method when Gaussian noise is added to the data with a signal-to-noise ratio (SNR) varying in $\{10, 20, 40\}$ dB.

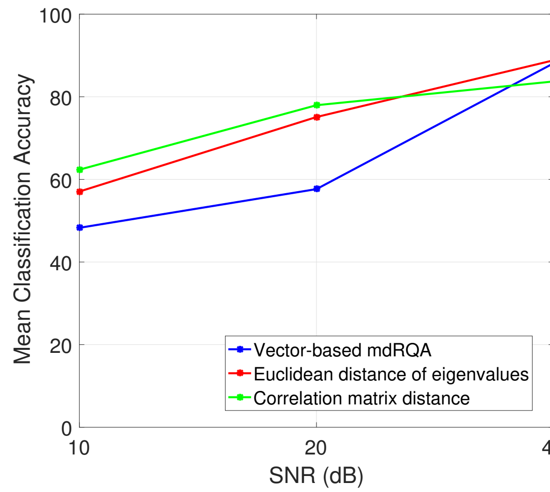

Fig. 3. Average classification accuracy of conventional mdRQA and our GmdRQA (for the Euclidean and CMD distance metrics), as a function of additive Gaussian noise's strength.

D. Evaluation Results

Table I displays the classification accuracy averaged over 100 repetitions, for the vector-based mdRQA approach and our proposed method. As it can be seen, the Euclidean distance between the eigenvalues vectors of state matrices outperforms the rest in terms of classification accuracy. Precision and recall percentages for each architecture examined are provided in Table II. Precision expresses the percentage of the results which are relevant, while recall refers to the percentage of total relevant results correctly classified by each algorithm.

Fig. 3 depicts the average classification accuracy as a function of the noise strength, for each one of the aforementioned architectures. For low SNR values, the correlation matrix distance outperforms the rest, whereas the vector-based mdRQA architecture presents the worst performance. On the other hand, for higher SNR values, the Euclidean distance between the eigenvalues vectors of state matrices presents the optimal performance, with the other two methods achieving a comparable accuracy. Overall, our GmdRQA method demonstrates an increased robustness to additive Gaussian noise, when compared against its vector-based mdRQA counterpart.

V. CONCLUSIONS AND FUTURE WORK

In this work, we designed and implemented a fully self-tuned architecture for the detection of dyslexia based on a representation of wearable sensor data in higher-dimensional phase spaces via a generalized mdRQA method based on state matrices for capturing the underlying dynamics of signal ensembles. The experimental evaluation on real data revealed the superiority of our proposed GmdRQA framework, when compared against its vector-based counterpart, in terms of classification accuracy and robustness to noise. As a future work, we intend to investigate the use of alternative distance measures and operators tailored to state matrices, in order to better capture specific characteristics of the dynamical system under study. Furthermore, we will extend our matrix-based GmdRQA to a more generic tensor-based framework, in order to model directly the inherent spatio-temporal dynamics of sensor streams.

References

[1] M. N. Benfatto, G. Ö. Seimyr, J. Ygge, T. Pansell, A. Rydberg, and C. Jacobson, "Screening for dyslexia using eye tracking during reading," *PloS one*, vol. 11, no. 12, p. e0165508, 2016.

[2] G. R. Lyon, S. E. Shaywitz, and B. A. Shaywitz, "A definition of dyslexia," *Annals of dyslexia*, vol. 53, no. 1, pp. 1–14, 2003.

[3] E. Temple, G. K. Deutsch, R. A. Poldrack, S. L. Miller, P. Tallal, M. M. Merzenich, and J. D. Gabrieli, "Neural deficits in children with dyslexia ameliorated by behavioral remediation: evidence from functional mri," *Proceedings of the National Academy of Sciences*, vol. 100, no. 5, pp. 2860–2865, 2003.

[4] N. Alexander-Passe, "How dyslexic teenagers cope: an investigation of self-esteem, coping and depression," *Dyslexia*, vol. 12, no. 4, pp. 256–275, 2006.

[5] L. Long, S. MacBlain, and M. MacBlain, "Supporting students with dyslexia at the secondary level: An emotional model of literacy," *Journal of Adolescent & Adult Literacy*, vol. 51, no. 2, pp. 124–134, 2007.

[6] K. Rayner, "Eye movements in reading and information processing: 20 years of research." *Psychological bulletin*, vol. 124, no. 3, p. 372, 1998.

[7] J. A. Kirkby, L. A. Webster, H. I. Blythe, and S. P. Liversedge, "Binocular coordination during reading and non-reading tasks." *Psychological bulletin*, vol. 134, no. 5, p. 742, 2008.

[8] F. Hutzler and H. Wimmer, "Eye movements of dyslexic children when reading in a regular orthography," *Brain and language*, vol. 89, no. 1, pp. 235–242, 2004.

[9] M. Biscaldi, S. Gezeck, and V. Stuhr, "Poor saccadic control correlates with dyslexia," *Neuropsychologia*, vol. 36, no. 11, pp. 1189–1202, 1998.

[10] M. H. Schneps, J. M. Thomson, G. Sonnert, M. Pomplun, C. Chen, and A. Heffner-Wong, "Shorter lines facilitate reading in those who struggle," *PloS one*, vol. 8, no. 8, p. e71161, 2013.

[11] I. Smyrnakis, V. Andreadakis, V. Selimis, M. Kalaitzakis, T. Bachourou, G. Kaloutsakis, G. D. Kymionis, S. Smirnakis, and I. M. Aslanides, "Radar: A novel fast-screening method for reading difficulties with special focus on dyslexia," *PloS one*, vol. 12, no. 8, p. e0182597, 2017.

[12] M. Wijnants, F. Hasselman, R. Cox, A. Bosman, and G. Van Orden, "An interaction-dominant perspective on reading fluency and dyslexia," *Annals of dyslexia*, vol. 62, no. 2, pp. 100–119, 2012.

[13] J. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *EPL (Europhysics Letters)*, vol. 4, no. 9, p. 973, 1987.

[14] S. Wallot, A. Roepstorff, and D. Mønster, "Multidimensional recurrence quantification analysis (mdrqa) for the analysis of multidimensional time-series: A software implementation in matlab and its application to group-level data in joint action," *Frontiers in psychology*, vol. 7, p. 1835, 2016.

[15] I. Vlachos and D. Kugiumtzis, "State space reconstruction from multiple time series," in *Topics on Chaotic Systems: Selected Papers from Chaos 2008 International Conference*. World Scientific, 2009, pp. 378–387.

[16] M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical review A*, vol. 45, no. 6, p. 3403, 1992.

[17] K. H. Krämer, R. V. Donner, J. Heitzig, and N. Marwan, "Dimension-scalable recurrence threshold estimation," *arXiv preprint arXiv:1802.01605*, 2018.

[18] J. Zbilut and C. Webber, "Embeddings and delays as derived from quantification of recurrence plots," *Phys. Lett. A*, vol. 171, pp. 199–203, 1992.

[19] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," in *Linear Algebra*. Springer, 1971, pp. 134–151.

[20] M. Herdin, N. Czink, H. Ozcelik, and E. Bonek, "Correlation matrix distance, a meaningful measure for evaluation of non-stationary mimo channels," in *2005 IEEE 61st Vehicular Technology Conference*, vol. 1. IEEE, 2005, pp. 136–140.