34th Annual International Conference of the IEEE EMBS
San Diego, California USA, 28 August - 1 September, 2012

An Association Framework to Analyze Dependence Structure in Time Series[†]

Bilal H. Fadlallah[1], Austin J. Brockmeier[1], Sohan Seth[2], Lin Li[1], Andreas Keil[3] and José C. Príncipe[1]

Abstract— **The purpose of this paper is two-fold: first, to propose a modification to the generalized measure of association (GMA) framework that reduces the effect of temporal structure in time series; second, to assess the reliability of using association methods to capture dependence between pairs of EEG channels using their time series or envelopes. To achieve the first goal, the GMA algorithm was updated so as to minimize the effect of the correlation inherent in the time structure. The reliability of the modified scheme was then assessed on both synthetic and real data. Synthetic data was generated from a Clayton copula, for which null hypotheses of uncorrelatedness were constructed for the signal. The signal was processed such that the envelope emulated important characteristics of experimental EEG data. Results show that the modified GMA procedure can capture pairwise dependence between generated signals as well as their envelopes with good statistical power. Furthermore, applying GMA and Kendall's tau to quantify dependence using the extracted envelopes of processed EEG data concords with previous findings using the signal itself.**

I. INTRODUCTION

Measures of dependence have been suggested in the literature to quantify dependencies in neural data. Examples include correlation [1], [2], mutual information [3], [4] and Granger causality [5], [6]. Recently, the generalized measure of association or GMA [7], [8], [9] has been applied on EEG time series to extract relational information between different recordings.

In this paper, we propose to further improve the performance of GMA when applied on time series and assess the relevance of applying the updated method on EEG. The incentive behind this is to lessen the effect of dependence induced purely by temporal structure. In fact, the reliance of GMA on nearest-neighbor computation reduces the convenience of using it with time-indexed series. The method designed to circumvent this limitation will be first validated on synthetically constructed time-series data for which the dependence level can be controlled. For this purpose, we start with a Clayton copula with a pre-determined value of Kendall's correlation that governs the dependence in the resulting random variables. The generated independent and identically distributed random variables are then smoothed, filtered, and enriched by noise to emulate EEG time-series data. The idea is to check whether GMA and Kendall's correlation are able to reject (accept) the null hypothesis of uncorrelatedness for a high (zero) pre-determined value of Kendall's correlation, when using either the raw signal or its envelope. Since the distribution of GMA under the null hypothesis is unavailable for GMA, we proceed by generating an empirical distribution for the null hypothesis.

The rest of the paper is organized as follows. In Section II, we briefly outline GMA and in Section III, we formulate an updated version of GMA that alleviates the impact of the temporal structure in realizations. Simulations using synthetic data are performed in Section IV. For this end, null hypotheses denoting uncorrelatedness were constructed and the statistical significance of the obtained results were assessed in view of each hypothesis. Simulations were also carried out on real EEG and compared to previous findings. Section V offers discussion and concluding remarks.

II. GENERALIZED MEASURE OF ASSOCIATION

It is known that correlation (in the sense of Pearson's coefficient) only captures second order interactions between any given time series. On the other hand, the selection of free parameters when estimating mutual information is a tedious problem. The motivation behind the generalized measure of association or GMA is to address these two concerns and hence exploit the benefits of capturing nonlinear structure without the cost of free parameters. Generally speaking, a measure of association estimates how often large values of a random variable are associated with large values of a second variable. GMA extends this idea by considering the pairwise distance between realizations instead of their values and computing a rank variable based on how relatively close the realizations are. The GMA value is finally computed as the skewness of this variable by calculating the area under the cumulative distribution function (CDF) of the rank variable. Alg. 1 describes the steps involved in computing GMA between two time series.

As measure of dependence, GMA is lower and upper bounded and invariant under rotation and scaling. The values it assumes range between 0.5 and 1 and it may be asymmetric. The fact that it is parameter-free gives it a unique computational advantage over other approaches. Typical sample sizes when computing GMA in practical applications would be greater than 50 samples.

III. TIME SERIES GMA

Typically, when considering realizations from two time series, the nearest neighbor in amplitude for a given point is

[†] This work was supported by the National Science Foundation under Grant IIS-0964197 and the Lebanese Center for Scientific Research (CNRS).

[1] B. Fadlallah, A. Brockmeier, L. Li and J. Príncipe are with the Department of Electrical and Computer Engineering, University of Florida, Gainesville, FL 32611, U.S.A. E-mail: {bhf, ajbrockmeier, linli, principe} at cnel.ufl.edu

[2] S. Seth is with the Department of Information and Computer Science, Helsinki Institute for Information Technology, Aalto University, Finland. E-mail: sohan.seth at hiit.fi

[3] A. Keil is with the Department of Psychology, University of Florida, Gainesville, FL 32611, U.S.A. E-mail: akeil at ufl.edu

978-1-4577-1787-1/12/$26.00 ©2012 IEEE

Algorithm 1: Generalized Measure of Association

```
Input: Bivariate time series $\{u_t, v_t\}_{t=1}^{n}$ assuming values in
    the joint space $\mathcal{U} \times \mathcal{V}$
Output: Estimated dependence $d \in [0.5 : 1]$
Initialization: $P(R = r) = 0 \; \forall \; r \in \{1, \ldots, (n-1)\}$
for $i \in \{1 \ldots n\}$ do
    - Find $u_{j*}$ ($j^* \in \mathcal{J}$) closest to $u_i$, equivalently
      $j^* = \underset{j \neq i}{\arg\min}\ \delta_u(u_i, u_j)$, where $\delta_u$ denotes
      Euclidean distance in $\mathcal{U}$.
    - For all $j^* \in \mathcal{J}$, find the spread of ranks, i.e.
      $r_{i,max}$ and $r_{i,min}$ of $v_{j*}$ in terms of $\delta_v$ such that:
        $r_{i,max} = \#\{j : j \neq i, \delta_v(v_j, v_i) \leqslant \delta_v(v_{j*}, v_i)\}$
        $r_{i,min} = \#\{j : j \neq i, \delta_v(v_j, v_i) < \delta_v(v_{j*}, v_i)\}$
    - For all rank values $r_{i,min} < r \leqslant r_{i,max}$, assign:
      $P(R = r) = P(R = r) + 1/|\ \mathcal{J}\ |/(r_{i,max} - r_{i,min})/n$
- Compute $C$ as the empirical CDF of $\{r_1, \ldots, r_n\}$.
  $d$ is the area under $C$ normalized by $(n-1)$
```

simply the nearest in time, however this does not reveal dependence structure. To overcome this obstacle, we propose to modify the GMA routine by decreasing the effect of temporal structure in the input time series. Again, the leading incentive behind this is that a pair of realizations from each time series will most probably be very close to the pair(s) corresponding to the closest in time. Therefore, for each realization in the time series, we dismiss the realizations within a neighboring time window to discard dependence purely pertaining to time structure. Only points falling outside that window would be considered as nearest neighbors in amplitudes. However, the choice of this window length is not a straightforward task. We suggest to use a window size intrinsic to the input domain and determined by the zero-crossing of the autocorrelation function (ACF) for each input time series. If no such crossing exists, we choose a lag corresponding to the first minimum of the ACF, or its $1/e$ decay if it does not achieve one. The aim is to decrease the correlation over time as much as possible and hence avoid misinterpreting high intrinsic association within each time series for high values of interdependence. This choice works well in our context although other choices are possible. The advantage brought by such setting is to keep the method parameter-free. The updated algorithm is outlined in Alg. 2 and Fig. 1 shows an illustrative example:

Algorithm 2: Time Series GMA (TGMA)

```
- Same Input, Output and Initialization as in Alg. 1
- Let $\xi_u$ and $\xi_v$ denote the ACFs of $u$ and $v$.
- Let $l_u^*$ be the lag at the first zero-crossing of $\xi_u$ if it exists,
  the first minimum of $\xi_u$ if no such crossing exists or the $1/e$
  decay level of $\xi_u$ if the latter is monotonically decreasing.
- Define $l_v^*$ similarly with respect to $\xi_v$ ($0 < l_u^*, l_v^* < n$).
for $i \in \{1 \ldots n\}$ do
    a. Find $u_{j*}$ ($j^* \in \mathcal{J}$), where $j^*$ satisfies:
           $j^* = \underset{|j-i| \geqslant \max(l_u^*, l_v^*)}{\arg\min}\ \delta_u(u_i, u_j)$
    b. For the obtained $j^*$, proceed as in Alg. 1 to compute
       the ranks spread and update $P(R = r)$.
- Return the area under the empirical CDF of the ranks.
```

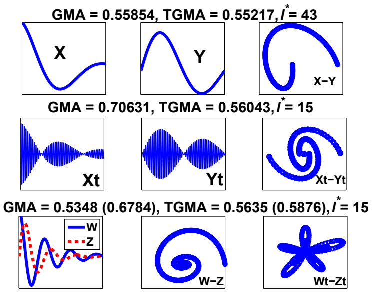


Fig. 1: GMA vs. TGMA and effect of temporal structure. **Top:** Two signals X and Y and their joint scatter plot. GMA and TGMA give close association estimates. **Middle:** Data is modulated using a carrier signal hence inducing significant temporal structure towards which TGMA shows less sensitivity. **Bottom**: Same procedure for two different signals.

IV. SYNTHETIC DATA

Since GMA is a rank-based approach, we select another rank-based method for comparison purposes. A good candidate is Kendall's rank correlation because of its simplicity and widespread use in a variety of applications.

A. Kendall's Rank Correlation

Kendall's rank correlation addresses some of Spearman's rank correlation (or Spearman's rho) insensitivities to special kinds of dependence. In contrast to Spearman's rho that proceeds with measuring the difference in the ranks of every pair of observations, Kendall's correlation measures in a non-parametric fashion the degree of association between two variables in terms of the number of occurrences of concordant N_c and discordant N_d pairs. For two time series $\{x_t, y_t\}_{t=1}^{n}$ of length n

1) N_c corresponds to cases where $\{x_i > x_j$ and $y_i > y_j\}$ or $\{x_i < x_j$ and $y_i < y_j\}$.
2) N_d corresponds to all other cases where $\{x_i > x_j$ and $y_i < y_j\}$ or $\{x_i < x_j$ and $y_i > y_j\}$.

The correlation measure generally referred to as Kendall's tau is then defined as:

$$\tau = (N_c - N_d) / \binom{n}{2}. \tag{1}$$

B. Input

We use two random vectors generated from the bivariate Clayton copula, with a scalar parameter $\tau_{desired}$. The Clayton copula is an asymmetric Archimedean copula, and can be defined as:

$$C_a(u, v) = \phi_a^{-1}(\phi_a(u) + \phi_a(v)) \tag{2}$$

where ϕ_a is the generator of the copula, a continuous and strictly decreasing convex function from $[0, 1]$ to $\mathbb{R}^+$, with $\phi_a(1) = 0$. An important property of an Archimedean copula is that it is directly related to Kendall's τ according to the following equation:

$$\tau_a = 1 + 4 \int_{t=0}^{t=1} \frac{\phi_a(t)}{\phi_a'(t)} \mathrm{d}t \tag{3}$$

Let C_θ be a Clayton copula. We have $C_\theta(u, v) = [\max(u^{-\theta} + v^{-\theta} - 1, 0)]^{-1/\theta}$, where the generator function is $\phi_\theta(t) = (t^{-\theta} - 1)/\theta$, hence:

$$\tau_{desired} = \tau_\theta = \theta / (\theta + 2) \tag{4}$$

We use a similar setting to the one in [7]. The length of the generated vectors was set to $114 \times 40 = 4560$ samples. Applying the inverse of a Beta CDF on the generated vectors generates two Beta random variables with the desired rank correlation ($\tau_{desired}$). The choice of a Beta distribution is to capture the amplitude constraints on EEG signals due to sampling [10], [11]. We choose the parameters α and β of the Beta distribution to be equal ($\alpha = \beta = 2$) resulting in symmetric distributions.

C. Synthetic Signal Processing

The generated random variables can be seen in Fig. 2. A Gaussian FIR filter with a 3-dB bandwidth-symbol time product of 0.16 is used to smooth the resulting signal. After subtracting the mean, we amplify the signals to the range of EEG data and add WGN with zero-mean and unit variance. A high pass filter with order 150 and a 12 Hz cutoff frequency and a lowpass filter with order 150 and 20 Hz cutoff frequency were used to bandpass the resulting signals to match the processing of the raw EEG data in [7].

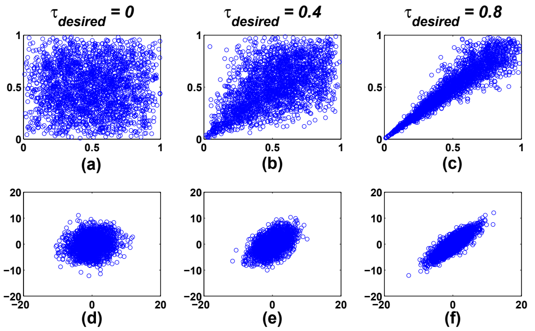


Fig. 2: Two random vectors generated from the bivariate Archimedean Clayton copula with a parameter corresponding to $\tau_{desired}$ of **(a)** 0, **(b)** 0.4 and **(c)** 0.8. **(d-e-f)** Corresponding Beta random variables.

Since we are interested in narrowband signals, we then extract the envelope as the instantaneous amplitude of the processed signal [12]. As a reminder, the instantaneous amplitude is defined as the magnitude of the analytic signal, in turn defined as the sum of the signal with its Hilbert transform. For a signal $w(t)$, the Hilbert transform $\mathcal{H}(w(t))$ can be expressed as:

$$\mathcal{H}(w(t)) = \frac{1}{\pi}\, w(t) * \frac{1}{t} = \frac{1}{\pi}\, PV \int_{-\infty}^{+\infty} \frac{w(t)}{t-\tau} \mathrm{d}\tau \qquad (5)$$

where PV denotes the Cauchy principal value of the singular integral. Fig. 3 shows the generated signals and their envelopes, versus real EEG data processed according to [7].

V. Simulations

A. Synthetic Data

Monte Carlo simulation consisting of 1000 iterations were performed to generate the null hypotheses of uncorrelatedness (induced by setting $\tau_{desired} = 0$) for synthetic data. Delays of up to 20 samples were introduced when computing dependencies between the signals to emulate propagation delays. Hence a total of 20000 points were used in the generation of the distribution corresponding to a given null hypothesis. Distributions corresponding to four null hypotheses were constructed: for each measure of dependence (GMA and Kendall's tau) two per type of input used (signal or envelope). The resulting distributions are shown in Fig. 4; the signals and their envelopes have similar distributions.

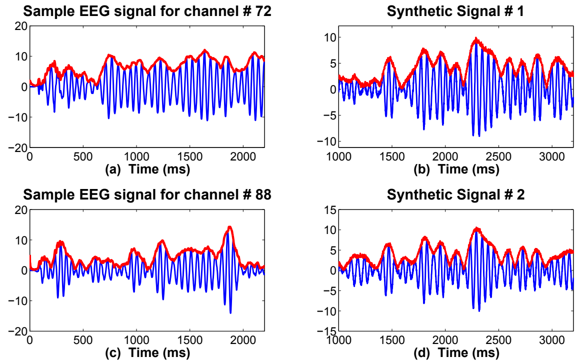


Fig. 3: **(a and c)** EEG signals recorded at channel locations 72 (a) and 88 (c) with their corresponding envelopes computed using the Hilbert transform. **(b and d)** Synthetic input generated as described in Section IV.B with the corresponding envelopes.

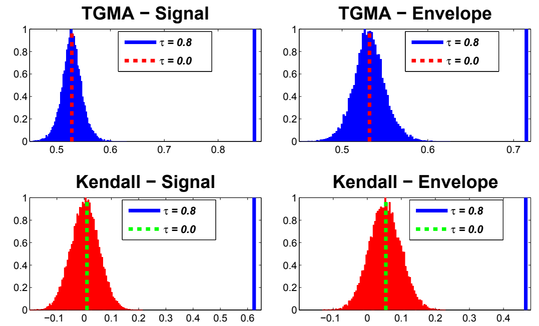


Fig. 4: Null hypotheses for uncorrelatedness generated at the signal and envelope levels when using TGMA for computations (upper row) and Kendall's tau (lower row).

We can then use these distributions to assess whether the observed envelopes of the processed signals are dependent as well. For example, the signals in Fig. 3 were synthesized using random variables correlated at a $\tau_{desired} = 0.8$ level. The computed correlation of the corresponding envelopes is 0.71 and obviously rejects the null hypothesis with a very small p-value. Hence, we show that the filtering, noise addition, and envelope extraction steps do not shield the intrinsic dependence value. Following this observation, we use the envelopes of the processed signals in the remainder of this manuscript.

B. Real Data

We propose to tackle the problem suggested in [7] by applying the methods mentioned above on the envelope signal. As a brief overview, the experimental setting exploits the steady-state visual evoked potential (ssVEP) paradigm by flashing a visual stimulus at a rate of 17.5 Hz to a participant. Two types of stimuli were presented to the subject, one representing an image of a neutral human face and the second a Gabor patch, each presented for a duration of 4.2 sec (plus 0.4 sec pre-stimulus baseline). A surface Laplacian method was applied on the raw EEG data and the experiment's goal was to identify which regions are active during the cognitive processing of each stimulus and hence analyze the corresponding connectivity patterns between all channel locations. In [7], two traditional coupling methods (Pearson's correlation and mutual information) besides GMA were used to calculate bivariate interactions with respect to a single parietal channel chosen as reference. The methodology suggested in this paper will be applied on the same experimental data to compare the inferred functional relationships across different electrode sites.

C. Results

Tests on synthetic data were repeated for values of $\tau_{desired}$ ranging from 0 to 1 and confirmed the observation of Fig. 4. This is reflected by the increasing values of captured dependence in Fig. 5 (a & b) and represents a motivation to extract dependency information from envelopes of processed time series without losing track of the underlying dependence.

Fig. 6 shows the obtained dependence maps when using Kendall's tau and Time Series GMA (TGMA) on EEG data for the face condition. Both measures indicate higher coupling for this condition between occipital sites and the temporal-parietal-occipital sites neighboring electrode P_4. Using the distributions in Fig. 7 (obtained via surrogate data generation by permutation of samples), those locations would correspond to regions exhibiting statistically significant dependence with respect to the reference electrode. The number of statistically significant pairwise links can be seen in Fig. 5 (c). The small significance level used is 0.039% and has been determined by Bonferroni's correction criterion for multiple comparisons where we divide the family wise error rate of 5% by the number of performed comparisons (129 in this case). With this method, TGMA has 51 significant links for the Face condition and 32 for the Gabor condition. For Kendall's tau, the numbers are respectively 90 and 72. The number of common links returned by both methods is 41, corresponding to 81% of the total TGMA links.

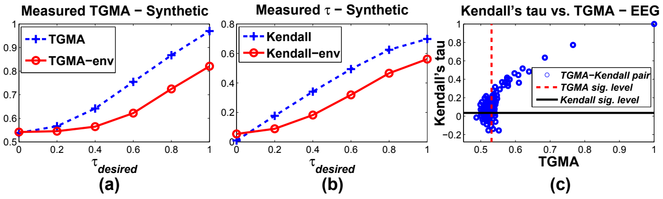
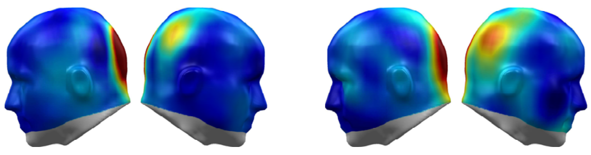


Fig. 5: **(a and b)** Measured values of TGMA and Kendall's tau versus desired dependence levels. **(c)** Obtained p-values with the corresponding statistical significance. We consider a test to be statistically significant when the p-value is less than 0.00039. Most measured τ values were significant.



(a) TGMA-L (b) TGMA-R (c) Kendall-L (d) Kendall-R

Fig. 6: First two subplots (a and b) show interpolated TGMA measures over right and left (R and L) head surface for the Face condition and subsequent subplots (c and d) exhibit the same when using Kendall's tau.

VI. CONCLUSION

This paper demonstrated on synthetically generated data that it is possible to apply measures of association on the envelopes of processed signals to quantify dependence between the underlying random variables. Based on this, we carry out a similar procedure on a single-trial EEG dataset where the goal is to localize functionally connected regions in response to a certain visual stimulus. The EEG time series are first processed as depicted in [7] and [8], then the instantaneous amplitude of the signal is extracted and pairwise association measures were computed using both Kendall's tau and TGMA. Results show that both measures seem to indicate activation in a region near P_4 with overlap in statistically significant links between the two measures. The large number of channels labeled by Kendall's tau as dependent with statistical significance suggests that the latter is more sensitive to transient dependencies and that TGMA performs better in capturing non-temporal dependence. Future work includes looking at the dependence evolution in time from a dynamic graph theoretical perspective.

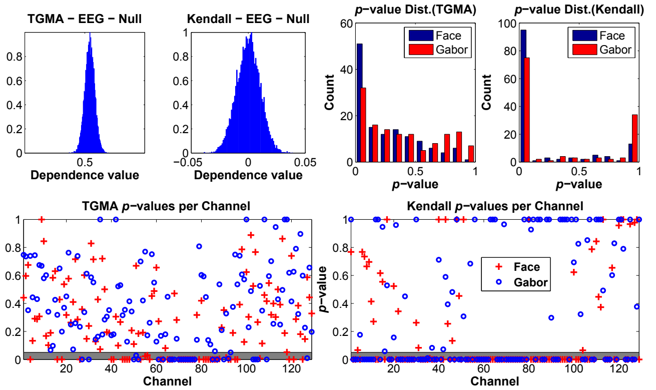


Fig. 7: Top row shows the generated null hypotheses using EEG data, and the distribution of the obtained corresponding p-values. Bottom row plots the obtained p-value per channel per condition for TGMA / Kendall's tau.

REFERENCES

[1] K. Schindler, H. Leung, C. E. Elger and K. Lehnertz, "Assessing seizure dynamics by analysing the correlation structure of multichannel intracranial eeg", in Brain, vol. 130, no. 1, pp. 65-77, 2007.

[2] M. Guevaraa and M. Corsi-Cabrera, "EEG coherence or EEG correlation?", in Int. J. Psychophysiol., vol. 23, no. 3, pp. 145-153, 1966.

[3] J. Jeong, J. Gore and B. Peterson, "Mutual information analysis of the EEG in patients with Alzheimer's disease", in Clin. Neurophysiol., vol. 112, no. 5, pp. 827-835, 2001.

[4] S. Na, S. Jin, S. Kim and B. Ham, "EEG in schizophrenic patients: mutual information analysis", in Clin. Neurophysiol., vol. 113, no. 12, pp. 1954-1960, 2002.

[5] W. Hesse, E. Moller, M. Arnold and B. Schack, "The use of time-variant EEG Granger causality for inspecting directed interdependencies of neural assemblies", in J. Neurosci. Methods, vol. 124, no. 1 pp. 27-44, 2003.

[6] M. Ding, Y. Chen and S. Bressler, "Granger Causality: Basic Theory and Application to Neuroscience", in Biol. Cybern., vol. 85, no. 2, pp. 145-157, 2006.

[7] B. Fadlallah, S. Seth, A. Keil and J. Principe, "Robust EEG preprocessing for dependence-based condition discrimination", in Proceedings of the 33rd International Conference of the EMBS, pp. 1407-1410, 2011.

[8] B. Fadlallah, S. Seth, A. Keil and J. Principe, "Analyzing dependence structure of the human brain in response to visual stimuli", in Proceedings of the 37th ICASSP, Kyoto, Japan, 2012.

[9] B. Fadlallah, S. Seth, A. Keil and J. Principe, "Quantifying cognitive state from EEG using dependence measures", in IEEE Trans. Biomed. Eng. (submitted), 2012.

[10] P. Peebles, Probability, Random Variables and Random Signal Principles, 4th edition. Singapore: McGraw Hill, 2001.

[11] N. Stevenson, M. Mesbah, G. Boylan, P. Colditz and B. Boashash, "A nonlinear model of newborn EEG with nonstationary inputs", in Ann. Biomed. Eng., vol. 38, no. 9, pp. 3010-3021, 2010.

[12] B. Boualem, "Estimating and interpreting the instantaneous frequency of a signal-part 1 : fundamentals", in Proceedings of the IEEE, vol. 80, no. 4, pp. 520-538, 1992.