For my parents, to whom I owe everything
ACKNOWLEDGMENTS

I shall take this opportunity to thank the people who have played a significant role in my Ph.D. journey. I am indebted to my advisor, Dr. José C. Príncipe, for his guidance, insights, encouragements and financial support during the past years. He is a true inspiration and it is difficult for me to keep track of the things I learned from him.

I have also been lucky to work closely with Dr. Andreas Keil, who introduced me to cognitive psychology and research in that field. I won’t forget his enthusiasm or patience with my stupid questions. I would like to thank Dr. Tan Wong and Dr. Dapeng Wu for being kind enough to serve on my committee, Dr. Vladimir Miskovic for helping me with collecting EEG data, Dr. Firas Kobaissy for being there all along, and all my colleagues at CNEL for being such great friends, especially those who have been on the journey since the beginning and the Football superstars in F.C. Principe. The fun moments, research and random discussions, and the after-work events will be hard to forget.

I am indebted to the Lebanese National Council for Scientific Research (CNRS), which was instrumental in providing financial support during my time at UF. I specifically thank Dr. Charles Tabet for his continuous encouragement and support and the many interesting and fruitful discussions. Special thanks also go to Erion Hansanbelliu, Rakesh Chalasani, Goktug Cinar, Miguel Teixeira and Gabriel Nallathambi for helping me with the EEG recordings.

Graduate years were very important to me in that they allowed me to better understand the fine balance that exists in research between the subjective human approach and the scientific reality. It was a challenging ride where I had the opportunity to attend many inspiring talks, read several interesting papers and foremost, learn the importance of questioning the basics and venturing into unexplored territories. I learned something from every CNELer and am thankful for that Sohan, Badong, Erion, Alex, Manu, Bot, Steve, Luis, Austin, Goktug, Evan, Rakesh, Songlin, Lin, Jihye, Ping, Injun, Matt, Rosha,
Gabriel, Gavin, Eduardo, Stefan and everyone who contributed to making CNEL a productive and healthy environment for research and discussions (fruitful or not-that-fruitful).

Finally, I would like to thank my family: my parents, two brothers and two sisters, for being the ultimate source of joy, strength, support, and inspiration in my life, and of course as always:

اَلَّلَهُمَّ}}
ACKNOWLEDGMENTS ................................................................. 4
LIST OF TABLES ................................................................. 10
LIST OF FIGURES ................................................................. 11
ABSTRACT ..................................................................................... 14

CHAPTER

1 INTRODUCTION ................................................................. 16

1.1 Background ................................................................. 16
  1.1.1 Brain Visualization Techniques ........................................ 16
  1.1.2 The Electroencephalogram (EEG) ....................................... 18
    1.1.2.1 History ............................................................. 18
    1.1.2.2 Neurophysiological concepts .................................... 18
    1.1.2.3 EEG versus MEG ................................................. 19
  1.1.3 EEG for Cognitive State Evaluation ................................... 20
  1.1.4 The Steady-State Visual Evoked Potential (ssVEP) ............... 21
  1.1.5 Face Perception and Attention Systems ............................... 21
  1.1.6 Problem Statement and Research Goals .............................. 24

1.2 Procedure Overview ............................................................ 24
  1.2.1 EEG Signal Processing Scales ........................................... 24
    1.2.1.1 Time domain methods ........................................... 24
    1.2.1.2 Frequency domain methods ................................... 25
    1.2.1.3 Time-frequency domain methods ................................. 25
    1.2.1.4 Nonlinear dynamics ............................................... 26
    1.2.1.5 Time scale choice .................................................. 27
  1.2.2 Signal Processing Approach ............................................. 27
  1.2.3 Measures of Dependence ................................................. 31
    1.2.3.1 Brain networks and functional connectivity ................... 31
    1.2.3.2 Why measures of dependence? .................................. 32
    1.2.3.3 Previous work ...................................................... 32
    1.2.3.4 Shortcomings in current methods ................................. 34
    1.2.3.5 Proposed methods ................................................. 36

1.3 Dissertation Organization .................................................. 37

2 EXPERIMENTAL SETUP AND SIGNAL PROCESSING .................. 38

2.1 Methodology ................................................................. 38
  2.1.1 Experimental Setting .................................................... 38
    2.1.1.1 Stimuli and recordings .......................................... 38
    2.1.1.2 Data .................................................................. 40
2.1.1.3 Channel subsets ........................................... 41
2.1.2 Signal Processing Overview ................................ 42
2.2 EEG Signal Processing ......................................... 44
  2.2.1 Raw Data ....................................................... 44
  2.2.2 Current Source Densities .................................... 45
  2.2.3 Signal Processing ............................................. 46
      2.2.3.1 Notch filters ........................................... 46
      2.2.3.2 Bandpassing ............................................ 47
      2.2.3.3 Testing .................................................. 48
2.3 Analysis in the Frequency-Domain ................................ 53
  2.3.1 Power ......................................................... 53
  2.3.2 Phase ........................................................ 54
3 MEASURES OF DEPENDENCE ......................................... 57
  3.1 Measures of Dependence in the Literature ..................... 57
      3.1.1 Dependence .................................................. 57
          3.1.1.1 Correlation ............................................ 57
          3.1.1.2 Mutual information .................................. 60
      3.1.2 Measures of Causality ...................................... 61
          3.1.2.1 What is causality? .................................... 61
          3.1.2.2 Cross-spectral analysis ................................ 62
          3.1.2.3 Granger causality ..................................... 64
      3.1.3 Measures of Synchrony ...................................... 66
          3.1.3.1 Phase-locking statistics ............................... 66
          3.1.3.2 Mean phase coherence .................................. 67
  3.2 Novel Measures of Dependence .................................. 68
      3.2.1 Preliminary Concepts ....................................... 68
          3.2.1.1 Dependence ............................................ 68
          3.2.1.2 Association ............................................. 69
          3.2.1.3 Generalized measure of association .................... 70
          3.2.1.4 Time series complexity and permutation entropy ... 72
      3.2.2 Generalized Association for Time Series .................. 75
          3.2.2.1 Motivation ............................................. 75
          3.2.2.2 Proposed algorithm ................................... 75
      3.2.3 Weighted-Permutation Entropy Based Dependence .......... 77
          3.2.3.1 Motivation ............................................. 77
          3.2.3.2 Weighted-permutation entropy .......................... 78
          3.2.3.3 WPE-based dependence .................................. 80
4 THE DYNAMIC DEPENDENCE GRAPH ................................ 81
  4.1 Graph Model .................................................... 81
  4.2 Graph Theoretical Concepts .................................... 82
      4.2.1 Basic Notations ............................................ 82
      4.2.2 Node Clustering Coefficient ............................... 82
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-1</td>
<td>Two-sample Kolmogorov-Smirnov test results for PE and WPE</td>
<td>105</td>
</tr>
<tr>
<td>6-2</td>
<td>Cronbach’s alpha per subject for different dependence measures &amp; conditions</td>
<td>118</td>
</tr>
<tr>
<td>6-3</td>
<td>Cronbach’s alpha across subjects, dependence measures and conditions</td>
<td>121</td>
</tr>
<tr>
<td>6-4</td>
<td>KS test results on averaged matrices for all subjects</td>
<td>123</td>
</tr>
<tr>
<td>6-5</td>
<td>Classification results for Subject 6.</td>
<td>123</td>
</tr>
<tr>
<td>D-1</td>
<td>Ratio of average entropies between epileptic &amp; non-epileptic segments</td>
<td>142</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-1</td>
<td>Spatial versus Temporal Resolution for different brain analysis tools.</td>
<td>17</td>
</tr>
<tr>
<td>1-2</td>
<td>First recorded human EEG from the first publication of Hans Berger.</td>
<td>19</td>
</tr>
<tr>
<td>1-3</td>
<td>Diagram visualizing face perception and attention systems.</td>
<td>23</td>
</tr>
<tr>
<td>1-4</td>
<td>Functional Brain Networks.</td>
<td>33</td>
</tr>
<tr>
<td>2-1</td>
<td>128-channel HCGSN montage used.</td>
<td>39</td>
</tr>
<tr>
<td>2-2</td>
<td>The two types of stimuli used to instigate ssVEPs.</td>
<td>40</td>
</tr>
<tr>
<td>2-3</td>
<td>Sensor map corresponding to the setting in Figure 2-1.</td>
<td>41</td>
</tr>
<tr>
<td>2-4</td>
<td>Sensor map corresponding to a 256-channels GSN montage.</td>
<td>42</td>
</tr>
<tr>
<td>2-5</td>
<td>The 10-20 system to describe the location of EEG electrodes.</td>
<td>43</td>
</tr>
<tr>
<td>2-6</td>
<td>Sensors in Figure 2-3 grouped according to a 10-20 scheme.</td>
<td>43</td>
</tr>
<tr>
<td>2-7</td>
<td>Standard parameters of a bandpass filter.</td>
<td>45</td>
</tr>
<tr>
<td>2-8</td>
<td>Original CSD signal at channel location 72 averaged over the 15 trials.</td>
<td>47</td>
</tr>
<tr>
<td>2-9</td>
<td>Change in number of channels where KS correctly discriminates conditions.</td>
<td>51</td>
</tr>
<tr>
<td>2-10</td>
<td>GMA vs. quality factor plot for 3 different filter orders.</td>
<td>52</td>
</tr>
<tr>
<td>2-11</td>
<td>Filtered CSD signal at channel location 72 averaged over the 15 trials</td>
<td>52</td>
</tr>
<tr>
<td>2-12</td>
<td>FFT power in electrode space for filtered signals averaged over trials.</td>
<td>53</td>
</tr>
<tr>
<td>2-13</td>
<td>Assessing power variability per trial at the flickering peak.</td>
<td>54</td>
</tr>
<tr>
<td>2-14</td>
<td>Power distributions for the Face and Gabor patch conditions.</td>
<td>55</td>
</tr>
<tr>
<td>2-15</td>
<td>FFT phase in electrode space for filtered signals averaged over trials.</td>
<td>55</td>
</tr>
<tr>
<td>2-16</td>
<td>Average power at $17.5 \pm 0.3$ Hz for each channel across subjects.</td>
<td>56</td>
</tr>
<tr>
<td>3-1</td>
<td>Simple illustration of GMA.</td>
<td>73</td>
</tr>
<tr>
<td>3-2</td>
<td>GMA vs. TGMA and effect of temporal structure.</td>
<td>77</td>
</tr>
<tr>
<td>3-3</td>
<td>Two examples of possible $m$-dimensional vectors sharing same motif.</td>
<td>78</td>
</tr>
<tr>
<td>4-1</td>
<td>An undirected graph consisting of three connected components.</td>
<td>84</td>
</tr>
<tr>
<td>4-2</td>
<td>Distributing the vertices of the dependence graph on a circle.</td>
<td>85</td>
</tr>
</tbody>
</table>
4-3 Effect of Free Parameters Selection. .................................................. 88
5-1 Computed values of $Q_{avg}^{FG}$ for TGMA, Kendall’s tau and Spearman’s rho. .. 96
5-2 Distributions of $Q_{m,n}^{FG}$ and $Q_{m,n}^F$ for TGMA, Kendall’s tau and Spearman’s rho. 97
6-1 Sensor network and channel 72 highlighted in red. ................................. 103
6-2 Dependence maps in sensor space. ......................................................... 104
6-3 Averaged phase-locking values for the face and Gabor patch conditions. ....... 106
6-4 Distributions of the averaged phase-locking values for the two conditions. .... 106
6-5 Using absolute correlation to weight graph connections for channel 72. ....... 107
6-6 Using GMA to weight graph connections for channel POz or 72. ............... 108
6-7 Using GMA to weight graph connections for the second subject. ............... 109
6-8 Using MI to weight graph connections for channel POz or 72. ................. 109
6-9 Electrode subset considered across time windows. ................................. 110
6-10 Using Permutation Entropy measures to weight graph connections for channel 72. 110
6-11 Empirical CDFs per condition for PE and WPE. ................................. 111
6-12 Using TGMA and Kendall’s tau to weight graph connections for channel 72. .. 111
6-13 Generated Null Hypotheses and Obtained p-values. ............................. 112
6-14 Obtained TGMA values versus Kendall’s tau. ..................................... 113
6-15 Dynamic dependence graph using TGMA (Face condition) ..................... 114
6-16 Dynamic dependence graph using TGMA (Gabor patch condition) .......... 115
6-17 Dynamic dependence graph using absolute correlation (Face condition) .... 116
6-18 Dynamic dependence graph using absolute correlation (Gabor patch condition) 117
6-19 Graph theoretical measures extracted from two graphs constructed using TGMA. 118
6-20 Graph theoretical measures from a graph constructed using Spearman’s rho ... 119
6-21 Betweenness centrality for a graph constructed using TGMA. ............... 119
6-22 Channels and their corresponding connected component sizes. .................. 120
6-23 Cardinality of each channel’s connected component mapped to sensor space. .. 121
6-24 Empirical cumulative distribution functions for the three dependence measures. 122
<table>
<thead>
<tr>
<th>No.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6-25</td>
<td>Test statistics and window size.</td>
<td>122</td>
</tr>
<tr>
<td>A-1</td>
<td>Potential map vs. CSD map of a visual evoked potential (VEP).</td>
<td>132</td>
</tr>
<tr>
<td>D-1</td>
<td>PE versus WPE in the case of an impulse.</td>
<td>136</td>
</tr>
<tr>
<td>D-2</td>
<td>Different entropy measures applied on a Gaussian-modulated sinusoidal train.</td>
<td>138</td>
</tr>
<tr>
<td>D-3</td>
<td>Normalized PE and WPE values for different SNR levels.</td>
<td>139</td>
</tr>
<tr>
<td>D-4</td>
<td>WPE-based analysis performed on filtered EEG data.</td>
<td>141</td>
</tr>
<tr>
<td>D-5</td>
<td>WPE-based analysis performed on a single channel processed EEG signal.</td>
<td>142</td>
</tr>
<tr>
<td>D-6</td>
<td>Different entropy-based measures applied on epileptic EEG.</td>
<td>143</td>
</tr>
</tbody>
</table>
The question of localizing cognition in the human brain is an old and difficult one. Particularly challenging is to understand the fascinating human ability to recognize and identify faces. The exquisite capacity to perceive facial features has been explained by the activity of neurons particularly responsive to faces, found in the fusiform gyrus and the anterior part of the superior temporal sulcus. This dissertation hypothesizes and demonstrates that it is possible to detect automatically the recognition of faces solely from processed electroencephalograms (EEG) in an online fashion with high temporal resolution using measures of statistical dependence applied on steady-state visual evoked potentials (ssVEPs).

EEG recordings are first modeled as an indexed family of random variables belonging to a stochastic process. Measures of dependence exploit bivariate distributions among pairwise channel recordings, and is a more realistic approach to quantify the joint spatio-temporal data distribution than previous methods just working with the marginal distributions, since the latter implicitly assume statistical independence between measurements. Standard and novel dependence measures were applied to estimate dependence within the filtered current source density (CSD) data. Based on previous and recent literature, the analysis included measures of (i) linear and monotone correlation (Pearson’s $r$, Spearman’s $\rho$ and Kendall’s $\tau$), (ii) synchrony (using phase-locking statistics), (iii) mutual information (using $k$-nearest neighbors), and (iv) entropy (permutation and approximate
Novel approaches to quantify dependence are proposed using the concepts of generalized association (GMA and TGMA) and weighted-permutation entropy (WPE).

Dependencies between channel locations were assessed for two separate conditions elicited by distinct pictures flickering at a rate of 17.5 Hz. Filter settings were chosen to minimize the distortion produced by bandpassing parameters on dependence estimation. A dynamic graph visualizing the dependence evolution in time was generated for each condition and dependence measure. Several concepts from graph theory were adapted to analyze the resulting graphs and identify the active recording sites. Measures of centrality were particularly useful in determining the main channels involved in the cognitive response and a connected components analysis was employed to study in depth the network structure.

A classification framework based on information theoretical concepts is further developed by computing a similarity measure between two matrices storing the dependence information. This measure is then used to determine whether or not two matrices share the same condition. Besides, statistical analysis was performed for automated stimuli classification on six participants using the Kolmogorov-Smirnov test. Results show active regions in the occipito-parietal part of the brain for both conditions with a greater dependency between occipital and inferotemporal sites for the face stimulus. This aligns with previous evidence suggesting re-entrant organization of the ventral visual system, showing heightened re-entry when viewing meaningful or salient stimuli. Further research should investigate whether the communication pattern observed in this study is direct or enabled via one or more intermediate sources.
CHAPTER 1
INTRODUCTION

The human brain is one of the most complex self-organizing systems known. It roughly consists of 100 billion neurons, each having on average 7000 synaptic connections with its surrounding neurons. The enormous structural connectivity of this intricate network complicates the complete understanding of its various processes, and makes measuring the functioning human brain one of the most formidable endeavors ever undertaken in science or engineering [Gevins, 1984].

1.1 Background

1.1.1 Brain Visualization Techniques

The question of localizing cognition in the human brain is an old and difficult one [Posner et al., 1988]. Current analyses of the operations involved in cognition are relying on novel and powerful brain imaging techniques during cognitive tasks. Today, the most widespread used neuroimaging tool in the clinical and research worlds is functional magnetic resonance imaging (fMRI). fMRI is based on nuclear magnetic resonance and is a hemodynamic technique, i.e. it investigates neural activity by tracking changes in blood flow. Another hemodynamic technique, positron emission tomography (PET), is gaining more acceptance as a nuclear medicine tool useful in the context of evaluating tumors noninvasively [Otte and Halsband, 2006]. The main drawback of using these techniques is that their corresponding temporal resolution is bound to the hemodynamic response of neural activity and hence suffers grievously from the slow response induced by the latter. On the other hand, they have an excellent spatial resolution, not exceeding a few millimeters.

The need for other alternatives to analyze neural activity within very short time scales motivates the usage of electromagnetic techniques. Such techniques have been under development for more than a century now and circumvent the latency and intrinsic low temporal resolution induced by the blood stream in the human body, by recording
electrical or magnetic information simultaneously from multiple locations on the human scalp. Electromagnetic tools can be electroencephalographic (EEG) or magnetoencephalographic (MEG) and have the advantage of providing an excellent temporal resolution (a few milliseconds) but usually poor spatial resolution (a few centimeters). Given that communication between neurons and neural ensembles typically ranges in time scales between 1 ms and 100 ms depending on the neurons characteristics [Anderson, 2004], fMRI or PET are not well suited to assess cognitive tasks where time scales of interest do not exceed 100 ms. In this case, EEG and MEG represent a precious tool to track the dynamics of neural activity with high temporal resolution and hence gather a better understanding of the temporal neural correlates of cognitive processes [Cabeza and Nyberg, 2000]. Figure 1-1 shows the spatial and temporal resolutions of different tools used for neural information extraction.

Figure 1-1. Spatial versus Temporal Resolution for different brain analysis tools.

---

1 This figure has been reprinted from Parasuraman and Rizzo [2008] with slight modifications.
1.1.2 The Electroencephalogram (EEG)

1.1.2.1 History

The electroencephalogram (EEG) is a noninvasive record of brain electric potential from the scalp using one, or an array of, electrodes. The basics of electrophysiology can be traced to the contributions Galvani, Volta, Ohm and Faraday made to understanding electrical potential and electrical current. These concepts were well established by the middle of the nineteenth century [Collura, 1993], which allowed Richard Caton, as early as 1875 [Caton, 1875], to record electrical activity from the brains of rabbits and monkeys using a mirror galvanometer. It was not until 1929, though, that Hans Berger [Berger, 1929] performed the first recording of a human EEG using a Siemens double-coil galvanometer (Figure 1-2). Ever since, the EEG discipline has known a continuous evolution and has been combined with other components in multi-modal approaches to improve the level of knowledge about brain activity. For example, combining EEG recordings with fMRI imaging has been suggested by Huang-Hellinger et al. [1995], Babiloni et al. [2005], Gotman et al. [2006], and others. Such method however introduces residual artifacts in the EEG, that is mainly caused by the cardioballistogram (BCG) and the changing fields applied during the fMRI image acquisition [Allen et al., 2000], thus making the process of analyzing the resulting signals even a harder task.

1.1.2.2 Neurophysiological concepts

The human cerebral cortex is a folded structure, 2 to 5 mm thick and containing around $10^{10}$ neurons [Nunez and Srinivasan, 2006], whose activation causes the flow of local currents. EEG measures mostly the currents flowing as a result of dendritic excitations of pyramidal cells. These currents essentially originate from flows of ions (Na+, K+, Ca++, and Cl-) in neuron membranes [Atwood and MacKay, 1989]. Electric potential differences are then caused by summed postsynaptic potentials creating electric dipoles between soma (neuron cell bodies) and apical dendrites (neural cell branches).
The numerous types of synapses and the variety of neurotransmitters make the general picture more sophisticated [Teplan, 2002]. Large populations of neurons are needed to generate detectable electric signals on the head surface. Moreover, given the multiple layers between the scalp and neurons generating current flow (like the skin and the skull), the recorded signal detected by scalp electrodes is weak and has to be massively amplified [Teplan, 2002]. The elaborate generation of EEG from the neurons is beyond the scope of this dissertation. Further details can be found in Da Silva [1991] and Buzsaki et al. [2003]. In Appendix A, we also outline the basic underlying principles that govern the EEG neurophysics because of their relevance to some stages of this work.

1.1.2.3 EEG versus MEG

As a byproduct of EEG research, MEG technology started developing in the late 1960s [Cohen, 1968], and made rapid advances. While MEG and EEG share important characteristics, they differ drastically in other aspects. MEG is more expensive to set up and requires a dedicated laboratory with magnetic shielding. In an EEG setting,
recording sensors are attached to the participant’s scalp and the latter is allowed relative mobility, whereas in an MEG setting, participants must remain still and sensors are set in helmet-shaped dewars. Besides, recorded EEG signals have higher orders of magnitude than their MEG counterparts \((mV\text{ versus } fT)\) [Economides and Getov, 2012]. In this dissertation, and since temporal resolution is critical for an accurate assessment of the human brain activity, we rely on EEG to analyze the brain response to specific cognitive tasks. In particular, we resort to dense array EEG for that purpose. The latter is a method employed to record EEG with a number of electrodes far exceeding the usual number of electrodes utilized with standard techniques (approximately 20 electrodes). The motivation for this approach is to increase the spatial resolution of scalp EEG [Holmes, 2008].

1.1.3 EEG for Cognitive State Evaluation

Scalp electroencephalography has been widely used for assessing cognitive functioning and evaluating different cognitive states. Examples include studying cognitive impairment in presenile dementia [Johannesson et al., 1979], gender differences in cognitive abilities [Corsi-Cabrera et al., 1989], and the performance of cognitive and memory ability [Klimesch, 1999]. EEG has also been used to derive models for the simulation of cortical activity during cognitive tasks [Zavaglia et al., 2006] and analyze cognitive information processing in schizophrenic patients [Kirsch et al., 2000].

The improved understanding of the brain response to oscillatory stimuli enabled the design of a class of experiments where the impact of noise and artifacts is less important in the recorded EEG. These experiments exploit the fact that stimuli oscillating at a given rate induce an oscillatory response with the same frequency (or multiples of it). These induced signals are referred to as steady-state visual evoked potentials (ssVEPs) and discussed in more details in the next section.
1.1.4 The Steady-State Visual Evoked Potential (ssVEP)

Scalp-recorded EEG signals are inherently noisy, and therefore, we rely on the steady-state visual evoked potential (ssVEP) for two main reasons: first, its usefulness in electrocortical investigations of attentional processes over time [Moratti et al., 2004; Muller and Hillyard, 2000]; second, the ease with which they can be elicited. ssVEPs are continuous brain responses caused by flashing visual stimuli, generally modulated in intensity with a fixed rate usually less than 30 Hz and not smaller than 3 Hz. These scalp potentials can be captured as signals oscillating with a fundamental frequency equal to the stimuli flashing rate. ssVEPs have recently become a popular tool in cognitive and clinical neuroscience [Keil et al., 2003; Keil and Heim, 2009; Keil et al., 2005; 2009; Moratti et al., 2004; Morgan et al., 1996; Muller et al., 2006] and brain-machine interfaces (BMIs) or brain-computer interfaces (BCIs) [Cheng et al., 2002; Materka et al., 2007; Regan, 1979; Wang et al., 2006; Zhang et al., 2010]. Paradigms using ssVEPs have been discussed by Viallette et al. in a recent comprehensive review paper [Vialatte et al., 2010]. ssVEPs are easy to induce and have several other advantages like good signal-to-noise (SNR) ratio, relative immunity to artifacts and cost effectiveness. In addition, they require minimal subject training, and simplify feature extraction and analysis in frequency-space. Before outlining how ssVEPs are used in our experimental paradigm, we first review how attention is involved in detecting faces and recognizing facial identities.

1.1.5 Face Perception and Attention Systems

Now that we provided brief overviews of the main components of our framework, we move to describe its cognitive and neuroanatomic aspects. Of most concern here is how humans perceive different visual stimuli. Faces for example can be considered to be the most biologically and socially significant visual stimuli for humans [Palermo and Rhodes, 2007]. Most people tend to recognize birds as birds, tigers as tigers and penguins as penguins, i.e. automatically assign the visual stimuli to a class of objects. This remains true for other things we might perceive such as tables, chairs, cars and buildings, but
not for human faces. We always tend to perceive a human face as someone’s face. The brain processes involved in recognizing faces are already in place at age two and get fully developed by age 10. Still, children make more mistakes than adults in face recognition, which suggests that time is an important factor in learning how to recognize faces well [SocietyForNeuroscience, 2010]. The core system for face perception consists of three main components: the fusiform gyrus, the superior temporal sulcus and the inferior occipital gyri [Haxby et al., 2000].

A human face can be seen as a salient emotional stimulus, regardless of its expression, thus allowing discriminating family members from total strangers while at the same time conveying important information such as gender, race, direction of eye gaze etc. That being said, all faces, even those that can be labeled as neutral or bearing no particular expression can be thought of having emotional significance and probably privileged access to visual attention resources [Palermo and Rhodes, 2007]. Figure 1-3 illustrates the main constituents of the face perception and attention systems.

Event-related potential (ERP) results suggest that faces are categorized around 100 ms after stimulus, much earlier than the 200 ms required to categorize objects and words [Palermo and Rhodes, 2007; Pegna et al., 2004], which suggests that detecting faces is fast and efficient. Moreover, we know that the amygdala is involved in responding to all facial stimuli. However the response is more accentuated for emotional rather than neutral faces [Streit et al., 2003]. Extensive research has been performed to determine whether face processing is mandatory [Lavie et al., 2003; O’Craven et al., 1999], understand if faces are registered without conscious awareness [Stone and Valentine, 2003; 2004; Stone et al., 2001], study selective attention to faces [Corbetta and Shulman, 2002; Kastner and Ungerleider, 2001], evaluate the attentional resource allocation for face processing [Huang

---

2 The figure has been reprinted from Palermo and Rhodes [2007] with minor modifications.
Figure 1-3. Diagram visualizing face perception and attention systems. Colored rectangles represent the core components for face perception. **Yellow** refers to identity and semantic information processing. **Red** refers to emotion analysis. **Blue** refers to the fronto-parietal cortical network involved in spatial attention [Hopfinger et al., 2000]. **Solid** (dashed) lines indicate cortical (subcortical) pathways for rapid emotional expression processing. Multiple feedback connections are thought to exist between the different blocks.

and Pashler, 2005; Maurer et al., 2002], and analyze if a bias exists for attending faces rather than other types of objects [Johnson et al., 1991; Ro et al., 2001].

Haxby et al. [1999] further show that ventral extrastriate regions get activated besides the superior temporal sulcus and inferior and mid occipital gyri regions upon showing an inverted face for a subject. No difference however was witnessed when inverting the image of a house. A detailed and comprehensive list of studies can be found in the paper by Palermo and Rhodes [2007].
1.1.6 Problem Statement and Research Goals

We hence propose to design an end-to-end framework that processes EEG from the stage of signal extraction to that of feature analysis or classification in order to: (1) discriminate across different time scales two cognitive tasks performed by different subjects and (2) analyze for each condition the different functional connectivity patterns involved between sensors. To do so, we adopt a comparative approach based on the use of traditional and novel time-based measures of dependence, using at the same time a dynamic graph formulation of the examined functional networks.

1.2 Procedure Overview

1.2.1 EEG Signal Processing Scales

Most EEG analysis measures can be classified according to their scales of time and frequency. In the following, we review these scales and describe common ways of extracting meaningful information that can link different EEG recordings. Such recordings may be taken from the same participant at different trials, points in time or space or from different participants.

Given the complexity of cortical activity across recording sites, one important step is to quantify the functional connectivity among scalp locations and, ultimately, brain areas. This can be done in several ways, which are usually categorized as time domain [He et al., 2007; Hjorth, 1970; Keil et al., 2009; Klein and Davis, 1981; Maynard, 1979; Saltzberg and Burch, 1971; Vidaurre et al., 2009], frequency domain [Baccala and Sameshima, 2001; Makeig, 1996; Nunez and Srinivasan, 2006; Rappelsberger and Petsche, 1988], time-frequency domain [Blanco et al., 1995; Koenig et al., 2004; 2005] and nonlinear dynamics [Mayer-Kress, 1986; Stam, 2005] methods. In Sections 1.2.1.1, 1.2.1.2, 1.2.1.3 and 1.2.1.4, we give a brief overview of each scale.

1.2.1.1 Time domain methods

Time-domain methods have several advantages that make them appealing in time-varying or nonlinear systems. They allow using short data windows, truncating the input
waveforms to come up with solutions corresponding to specific time durations, besides more accurately studying the effect of non-stationary artifacts. A time domain method preserves the time resolution perfectly as compared with frequency or time-frequency methods. Time domain methods have been applied by Sutter [1992] and Farwell and Donchin [1988] in the context of designing brain machine interfaces (BMIs), by Saltzberg and Burch [1971] to estimate period analytic descriptors and Keil et al. [2009] to examine functional relationships among structures involved in emotional perception.

1.2.1.2 Frequency domain methods

The Fast Fourier Transform [Cooley and Tukey, 1965] is a clever computational technique based on the important concepts introduced by Fourier [1807] stating that “any continuous periodic signal could be represented by the sum of properly chosen sinusoidal waves”. Together with spectral analysis, it can be seen as one of the main frequency-based methods to analyze EEG. The results of EEG spectral analysis can be often grouped into the traditional frequency bands:

- δ when \( f < 4 \) Hz
- θ when \( 4 < f < 7 \) Hz
- α when \( 8 < f < 13 \) Hz
- β when \( 14 < f < 30 \) Hz
- γ when \( f > 30 \) Hz

Coherence-based measures of dependence are introduced and further discussed in Chapter 3. Further analysis can be found in Gevins and Rémond [1987] and Da Silva et al. [1986].

1.2.1.3 Time-frequency domain methods

“EEG time-frequency analysis” refers to all approaches that decompose EEG signals into magnitude and phase information at each frequency, and characterize their changes over time [Roach and Mathalon, 2008]. These approaches are different as they capture
different aspects of EEG magnitude and phase relationships. An overview and discussion of EEG time-frequency analysis is given in Roach and Mathalon [2008].


1.2.1.4 Nonlinear dynamics

The EEG is a nonlinear signal [Elbert et al., 1994]. The progress in the theory of nonlinear dynamical systems or “chaos theory” has reached a stage that allows its application to neural data [Başar and Bullock, 1990; Mayer-Kress, 1986; West, 1993] and its usage to study self-organization and pattern formation in brain networks [Stam, 2005].

Nonlinear EEG Analysis: Nonlinear EEG analysis can be traced back to a 1985 paper by Rapp et al. [1985] reporting correlation on human sleep EEG. Since EEG is a time-varying signal, it might occur that nonlinear dynamics, a field that was first introduced by Huygens in 1669 [Huygens, 1669] and further established by Poincaré in 1889 [Poincaré, 1889], is a natural path to follow. Briefly, a dynamic system is a system that changes its state over time. Here, a state is determined by the values of the variables in consideration. If we have \( n \) variables for example, the state can be represented as a point in an \( n \)-dimensional space, commonly referred to as state-space or phase-space. A dynamic system is governed by a set of equations (usually coupled differential equations) describing how the state of the system changes over time. However, the starting point of any investigation in clinical neurophysiology is usually not a set of differential equations, but rather a set of observations in the form of an EEG or MEG record [Stam, 2005].
**Time-Delay Embedding:** The most important step in nonlinear analysis consists in reconstructing, starting from a certain number of time series, an attractor in the state space corresponding to the underlying system. Embedding a time series is equivalent to converting it into a sequence of vectors in an \( m \)-dimensional embedding space. For a sufficiently high embedding dimension \( m \), Takens [1981] has shown that the reconstructed attractor share the same dynamical properties as the true attractor. Time-delay embedding is the most widespread applied embedding technique in the context of time series. It consists of reconstructing, for each time series, \( m \)-dimensional vectors by considering \( m \) consecutive values of the time series as values for the \( m \) coordinates of the vector.

Specifically, if we consider the time series \( \{x_t\}_{t=1}^T \), the corresponding time-delay embedding representation would be: \( x_j^{m,\tau} = \{x_j, x_{j+\tau}, \ldots, x_{j+(m-1)\tau}\} \) for \( j = 1, 2, \ldots, T-(m-1)\tau \), where \( m \) and \( \tau \) denote respectively the embedding dimension and time delay. Setting the values of the \( m \) and \( \tau \) has been the subject of much debate [Cellucci et al., 2003; Kennel et al., 1992; Rosenstein et al., 1994] and is beyond the scope of this dissertation.

### 1.2.1.5 Time scale choice

In this study, we consider the first approach since it is fairly possible that a cognitive state is highly volatile, and therefore, it should be quantified only over a short period of time i.e. with high time resolution. This can be done preferably in the time domain, since frequency domain approaches sacrifice the time resolution in order to extract precise frequency information. Another motivation for selecting a time domain method is the flexibility the latter provides with the handling of time windows.

### 1.2.2 Signal Processing Approach

As seen in the previous section, a lot of knowledge has been accumulated over the past decades on understanding the brain processes and anatomical components involved in the perception and processing of facial stimuli. It is clear that it is the evolution in time of the interaction amongst neural masses that produces the cognitive state. Since the brain is a hierarchical system with physical extent, the type of sensing will restrict the
type of activity that will be collected, and its resolution in time and space (in the head). But no matter what the sensing modality is, one should understand that the information to quantify cognitive brain states is contained in the joint space time distribution of the stochastic signals being collected. Recording scalp EEG can be thought of as taking a snapshot of the functioning of these processes that is blurred by the volume conduction properties of the brain sampled in time and space. The first effect can be mitigated by transforming EEG into current source density (CSD) domain (Refer to Appendix A). Using dense-array EEG also helps address the space resolution concern. Therefore it can be claimed that the empirical joint distribution provided by the multivariate EEG time series contains maximal information to quantify the underlying cognitive states.

Unfortunately this joint time varying distribution is hardly ever utilized to quantify brain activity because of its huge dimensionality. In fact, a stochastic process is an index family of random variables, one for each time sample. Moreover, the spatial distribution of the EEG is a multivariate random variable of size given by the electrode grid (or the sources). This joint distribution to be quantified properly would require an enormous amount of data that is impossible to collect, therefore several assumptions are normally done.

Perhaps the most important is stationarity, which means that the joint moments over time are just a function of the lags. Ergodicity allows us to estimate statistical quantities by temporal averages which is also universally utilized. If we further assume Gaussianity in the statistics, then just first and second order moments are necessary to quantify the statistical behavior over time. As is well known, EEG is not stationary and probably not a Gaussian process either. However, experience has shown that locally (at approximately 1 second windows), EEG statistics do not change appreciably [Bullmore and Sporns, 2009]. Starting from this point, we model each recorded EEG signal as a realization of a stochastic process, where each EEG epoch corresponds to one observation of the stochastic process. If the number of channels is $n$, we end up with $n$ observations of the stochastic
process where each observation can be modeled as a single random variable, hence the resulting model consists of an indexed family or vector of real-valued random variables under assumptions of *stationarity* and *ergodicity*. Unstationarity in EEG originates both from external factors (unstable electrode contacts, eye blinks, muscle artefact, skin artefact and other types of artefact), as well as from within brain sources because of the on-off type of communication between neural masses. Denoting \( S \) our stochastic process, we can express it as:

\[
S = \{U_t, t \in I\} \quad \text{with} \quad I = \{1, \ldots, T\} \tag{1-1}
\]

where each \( U_t \) is a random variable defined on a probability space \( \{\Omega, \mathcal{F}, P\} \) corresponding to each time sample. A multivariate spatial distribution of size \( n \) can be then defined based on these random variables. Denoting \( U_t = \{X_i, i \in K\} \) with \( K = \{1, \ldots, n\} \), we express the cumulative distribution function in terms of the joint distribution:

\[
F(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n) \tag{1-2}
\]

An expression for the joint cumulative distribution would be:

\[
F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f_{X_1,\ldots,X_n}(u_1, \ldots, u_n) \, du_1 \cdots du_n \tag{1-3}
\]

We can then write each marginal distribution \( f_{X_i}(x_i) \) as:

\[
f_{X_i}(x_i) = \int \cdots \int f_{X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n}(x_1, \ldots, x_{i-1}, x_{i+1} \ldots, x_n) \, dx_1 \cdots dx_{i-1} dx_{i+1} \ldots dx_n \tag{1-4}
\]

The joint distribution is \( n \)-dimensional. Since \( n \) is at least 64 for dense-array EEG (we use 129 and 257 EEG settings in our case), working with the joint distribution is simply out of question. Most approaches used in the EEG analysis literature work on individual
channels. That translates into using the individual distributions underlying each channel. Such approach would be justified only if the joint distribution could be written as a product of the marginals, i.e. if the considered random variables were independent. Therefore, working with individual channels in the hope that marginal distributions would provide all the intrinsic information of the recordings is a very strong assumption and a major simplification, only valid under independence, which of course is also unrealistic.

To overcome this problem, we propose to work with the joint bivariate distributions of channels over time. For $n$ channels, there are $\binom{n}{2} = \frac{n(n-1)}{2}$ bivariate distributions corresponding to each distinct pair of channels. One can argue that looking at the problem from a bivariate perspective only exploits the information provided by the pairwise joint distributions and still does not exploit to the full extent the available information. Although this is true, we argue that considering the bivariate distributions approximates the analysis towards a better exploration of the available information. A useful way to exploit these bivariate distributions and estimate the coupling between the pairwise channels is using measures of statistical dependence, but these measures should quantify the 2D joint distribution, not some of its moments as cross correlation. Pairwise models offer a crucial simplifying assumption in that the number of pairs is quadratic in the number of elements, not exponential.

To elaborate more on how good we can describe neural systems based only on knowing how pairs of elements interact, we refer to the work of Roudi et al. [2009], who analyzed the performance of quantitative descriptors derived from pairwise analysis in the context of spike trains. The authors try to answer the question of whether the efficacy of pairwise models in approximating the true joint distribution for subsets of biological systems containing a small number of elements can be as well generalized to the whole system (that includes a much higher number of elements). A quantitative, generic and most importantly system-agnostic answer is provided in terms of a crossover point (determined by a specific number of elements), beyond which small systems can be said
to possess predictive power for large systems. Within this spirit, using a high number of elements in our pairwise analysis should tend to give a better descriptor for the whole system. This specific aspect will hence motivate our computational approach.

1.2.3 Measures of Dependence

1.2.3.1 Brain networks and functional connectivity

Previous work on cognition presents evidence that areas that are coactive during cognitive tasks can be as well interdependent i.e. functionally connected, forming a cognitive control network [Bressler, 1995; Cole and Schneider, 2007]. Based on this observation, it is only natural to explore the functional connectivity as a tool to quantify a cognitive state. In this dissertation, we study this particular aspect of cognitive neuroscience, with a strong focus on discriminating two cognitive states in statistical terms solely from the associated functional connectivity across brain regions. To elaborate, we specifically propose to study whether it is feasible, starting from recorded EEG, to distinguish the brain response to facial versus non-facial visual stimuli and outline the witnessed difference in activated regions.

This approach would be motivated, as previously discussed, by the fact that stimuli with higher object-based complexity or specific semantic content undergo facilitated visual cortical processing [Bradley et al., 2003; Keil et al., 2009; Palermo and Rhodes, 2007]. In spite of the fact that many aspects of the neural mechanisms controlling this facilitation remain unknown, much of the present evidence seems to point towards the amygdaloid complex and the parieto-frontal cortex as origins of re-entrant modulation into lower-tier visual areas, when perceiving biologically significant stimuli [Damasio, 1998; Davis, 1998]. According to this hypothesis, visual perception and attention to relevant stimuli involve communication between the occipital and frontal cortices mediated by subcortical structures [Baizer et al., 1993; Lang et al., 1997]. Todorov and Engell [2008] have also suggested that novel faces are automatically evaluated by the amygdala according to a general valence dimension causing the activation of a face responsive network in the
occipital and temporal cortices, whereas the contribution of the fusiform gyrus and superior temporal sulcus in processing facial structures has been studied in Iaria et al. [2008]; Sergent et al. [1992] and others. Therefore, we exploit the fact that an increased functional connectivity in and among these cognitive areas is a precise indication of the presence of facial visual stimulation than a standard object.

Functional brain networks can be explored using graph theory [Bullmore and Sporns, 2009; Sporns et al., 2004] through standard steps as illustrated in Figure 1-4.

1.2.3.2 Why measures of dependence?

As mentioned in the previous section, coming up with a model for functional connectivity should be based on a measure that estimates the coupling between different channels. No matter how complex the processing done on the recorded waveforms is, it is very hard to assess how these recordings relate to each other without a measure that estimates the coupling, and this is exactly the motivation for using measures of dependence. Measures of dependence represent a common way to extract meaningful information that can link different EEG recordings by assigning scalar values to represent the degree of coupling between different recordings. Such recordings may be taken from the same subject at different trials, points in time or space or from different subjects.

1.2.3.3 Previous work

No consensus exists on an optimal way to quantify coupling among EEG recordings. The underlying reason for this lies in the multifarious nature of coupling. The latter can be categorized as linear or nonlinear, symmetric or nonsymmetric, synchronous or asynchronous, transient or nontransient. This is why the EEG literature contains many approaches to quantify dependency. Linear measures of dependence mainly include cross-correlation and coherence approaches and have been discussed by Corsi-Cabrera et al.

---

3 This figure has been reprinted from Bullmore and Sporns [2009] with slight modifications.
Figure 1-4. Steps involved in modelling functional brain networks: (1) define a set of network nodes, (2) estimate a continuous measure of association between nodes, (3) produce a binary adjacency matrix or undirected graph and (4) calculate the network parameters of interest in this graphical model of a brain network.

Nonlinear correlation measures have further been investigated by Fernandes de Lima et al. [1990], Pijn et al. [1992] and Allen et al. [1992]. On the other hand, nonlinear approaches include mutual information (MI) that have been discussed by Roulston [1999], Na et al. [2002], Honey et al. [2007] and Jeong et al. [2001], measures of generalized [Arnhold et al., 1999; Montez et al., 2006; Schiff et al., 1996; Stam et al., 1999] and phase [Lachaux et al., 1999; Tass et al., 1998].
association and measures of synchrony [Cavanagh et al., 2009; Gruber et al., 2001; Lachaux et al., 1999].

Cognitive processing usually requires integration of information processed simultaneously in spatially distinct areas of the brain [Montez et al., 2006]. If we model by a function the influence two different areas of the brain exercise on each other, that function will be more likely consist of linear and nonlinear terms. Depending on which of the terms is dominating, measures of linear dependence may not be able to catch the concerned interdependencies among the signals. Hence nonlinear measures of dependence may be more suitable to characterize functional coupling. The performance achieved by these measures has been the topic of a comparative paper by Quian Quiroga et al. [2002], in which phase synchronization, cross correlation and the coherence function give qualitatively equivalent results, with more sensitivity observed for nonlinear measures. Concerning mutual information, the following conclusions were drawn: (i) the performance of MI was dependent on the estimator used and the embedding technique (the first factor to a greater extent), (ii) MI does not produce robust estimates of synchronization, especially when the amount of available data is limited and (iii) that the best MI estimator in terms of performance achieved is a $k$-nearest neighbors estimator [Nicolaou and Nasuto, 2005; Quian Quiroga et al., 2002].

1.2.3.4 Shortcomings in current methods

Correlation and coherence based methods: As discussed, these are measures of linear dependence, and it is a strong assumption that two brain regions communicate through a linear channel. These methods disregard any nonlinear coupling between different recordings.

Granger causality based methods: Granger causality is a method that can be used in time or frequency domains to establish the directionality of linear dependence relationships (more about it in Chapter 3). Granger causality can be formulated using a linear vector autoregressive (VAR) modeling of time series. The fundamental assumptions
of the latter are stability, stationarity, and gaussian distribution of the error term. EEG could be considered stationary on time intervals of 1 second as shown by Bernasconi and Konig [1999] but such assumption remains questionable depending on the context used. Moreover, numerical experiments conducted by Kammerdiner [2008] indicate that the stability condition of vector autoregressive model is often violated in application to the EEG data.

**Phase synchrony based methods:** As a pre-requisite to estimating phase synchrony between two given time series, the phase of the signals at the frequency of interest needs to be determined first. The most popular standard approaches for phase estimation are the Hilbert transform and the Morlet wavelet approach [Lachaux et al., 2002]. The disadvantage brought by the first is the assumption of narrowbandness in the signals under consideration, which might not always be fulfilled (if that assumption is broken, instantaneous quantities might now be well defined). As a reminder, the instantaneous phase is defined as the phase 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:

$$\mathcal{H}(w(t)) = \frac{1}{\pi} w(t) \ast \frac{1}{t} = \frac{1}{\pi} \text{P.V.} \int_{-\infty}^{+\infty} \frac{w(t)}{t - \tau} d\tau$$

(1–5)

where P.V. refers to the Cauchy principal value. Hence the estimated phase becomes:

$$\phi_w(t) = \tan^{-1} \frac{\mathcal{H}(w(t))}{w(t)}$$

(1–6)

On the other hand, the Morlet wavelet (also known as the Gabor function) introduces two parameters as reflected by Equation 1–7, the rate decay $\sigma$ and the center frequency $w$ and imposes a non-uniform time-frequency resolution tradeoff on the extracted phase estimates, i.e. high (low) time resolution at low (high) frequencies [Aviyente, 2009].
Mutual Information based methods: MI is a powerful method to capture nonlinear structure and a more rigorous quantification of association between two signals. However it has three main drawbacks:

- For all finite sequences, there is a systematic overestimation of the mutual information [Herzel et al., 1994].
- It is difficult to choose the best estimator for MI for a specific case, especially when the sample size is small and the dimensionality is high.
- Estimating mutual information involves at least one free parameter, which adds more uncertainty to the analysis.

1.2.3.5 Proposed methods

To tackle these problems, we mainly adopt a generalized association approach. Typical advantages of such method is that it is parameter-free, and that it provides an intuitive understanding of dependence in the context of realizations. These are important when we only have a few samples – a situation where traditional estimators become highly biased, thus loosing their meaning. We elaborate more about these concepts in Chapter 3.

Therefore, the methods described in this thesis will contribute in:

1. Introducing two novel measures of dependence to use with time series. The first is based on general association, can capture nonlinear association and has an estimator that does not involve free parameters. The second is based on an ordinal symbolic dynamics to measure time series complexity. Although its estimator involves one free parameter, we show it is a powerful method to infer relationships in the considered data. To the best of our knowledge, using complexity measures in the context of dependence has not been previously suggested in the literature. This is discussed in Section 3.2.
2. Approaching the above concepts from a graph theoretical perspective to analyze brain functional networks. Several graph theoretical metrics will be used to quantify the dynamics of connectivities across time. Chapter 4 contains further elaboration on the dynamic dependence graph and the main characteristics derived from it.

3. Classifying trials in a supervised fashion using information theoretical concepts applied on the dependence matrices. More details on the adopted approach are available in Chapter 5.

1.3 Dissertation Organization

The rest of the dissertation is organized as follows. In Chapter 2, we go over the experimental setting, describe in detail the signal processing approach, and analyze the frequency components of the obtained signals, processed according to a robust filtering scheme that reduces the sensitivity of the dependence backend to bandpassing parameters. In Chapter 3, we examine current approaches in the literature that can be used to analyze dependence between different electrode measurements and infer associations between the corresponding channels. Besides these measures, we also describe novel ordinal-based measures that can be used to quantify dependence among EEG channels. In Chapter 4, we outline the procedure followed starting from the extracted dependence values to the construction of the dynamic dependence graphs. Chapter 5 discusses the use of information theoretic concepts in a classification context to estimate whether two dependence matrices belong or not to the same condition. Chapter 6 provides an overview of results from the previously mentioned perspectives based on a standard formulation of the problem, and Chapter 7 offers discussion and future work perspectives.
CHAPTER 2
EXPERIMENTAL SETUP AND SIGNAL PROCESSING

2.1 Methodology

2.1.1 Experimental Setting

The experiment included six participants. All were male graduate students in their early to mid-twenties who gave written consent prior to taking part in the study. The subjects had normal vision and no family history of epilepsy and did not report any psychotherapeutic history. The procedure was approved by the institutional review board of the University of Florida. The primary goal of this dissertation is to concentrate on the methodological aspects of the research process and validate the designed algorithms with real human data, while being able to assess variability across different subjects. The chosen number of subjects was considered sufficient to adequately characterize the cognitive behavior among a relatively small but meaningful population.

2.1.1.1 Stimuli and recordings

Two data acquisition systems from Electric Geodesics Inc. [2007] were used to record the electrocortical activity of the 6 participants: (i) a 128-channel Hydro-Cell Geodesic Sensor Net (HCGSN) and (ii) a 256-channel GSN 200 v.2.1 net (GSN-200). Three subjects used the 128-channel montage and the others the 256-channel montage. After applying the nets and setting electrode impedances below 50 kΩ according to the manufacturer’s recommendation for the high-impedance amplifiers, the experimental session proceeded with continuous recordings using $C_z$ as recording reference. An illustration of one of the data acquisition systems used can be seen in Figure 2-1.¹

Epochs of 400 ms prior to stimulus onset and 4200 ms after onset were extracted for a total of 4.6 seconds recordings. Subsequent to the baseline segment, an image showing a neutral human face was presented to the subject on a 17” monitor with a vertical refresh

¹ This figure has been reprinted from page 31 of Electric Geodesics Inc. [2007].
rate of 70 Hz. The same procedure was repeated with a control stimulus showing a Gabor patch, i.e. a pattern of stripes, where patches are calculated on the fly using pixels of the face picture. Luminance for both pictures was set to vary from near-zero to 52 cd.m$^{-2}$ and both were matched for mean luminance (i.e. 9.7 cd.m$^{-2}$), average contrast (i.e. 50%), and mean spatial frequency (i.e. 4 cpd) to preclude systematic differences with respect to these parameters. Figure 2-2 shows the two types of stimuli used in the experiment.

The stimuli flickered at a frequency $F_o = 17.5$ Hz and a total of $N_t = 15$ or $N_t = 20$ trials were performed. The rhythmic stimulation rate was designed to evoke an oscillatory electrocortical response at the driving frequency (steady-state visual evoked potential). From this point onward, we refer to the facial stimulus as “Face” and the Gabor stimulus as “Gabor patch”. During the experiment, each subject was asked to fixate the visual
Figure 2-2. The two types of stimuli used to instigate ssVEPs. Both were matched for mean luminance, average contrast and mean spatial frequency.

stimuli and attentively maintain gaze on the pictures, not to blink and avoid as much as possible movements of the eye and the head. EEG data was collected from the recordings of the $N_c = 129$ or $N_c = 257$ electrodes, at a digitization rate of $F_s = 1000$ Hz. In a typical scenario, we would perform offline artifact rejection (that might include some visual inspection steps) to detect individual channel artifacts and interpolate potentials at these channels via spherical splines. This was not performed in our case since it was not clear how it would affect computed dependencies in sensor space.

2.1.1.2 Data

The recording at each channel location $i$ was collected as a time series $x^{(i,k)}$ for face and $y^{(i,k)}$ for Gabor patch per trial $k$. As a result, the whole data can be represented as two spatio-temporal data matrices $X^{(k)}$ and $Y^{(k)} \in \mathbb{R}^{N_c \times N_s}$:

$$X^{(k)} = \begin{bmatrix} x^{(1,k)} \\ \vdots \\ x^{(N_c,k)} \end{bmatrix} \quad \text{and} \quad Y^{(k)} = \begin{bmatrix} y^{(1,k)} \\ \vdots \\ y^{(N_c,k)} \end{bmatrix},$$
where $N_c$ and $N_s$ denote the number of channels and sampled data points per recording, respectively. $N_c = 129$ for the 128-channel HCGSN and $N_c = 257$ for the 256-channel GSN-200 network (this is since the reference electrode is also counted). For $F_s = 1000$ and a downsampling factor $d = 1$, $N_s = 4600$. For later use, we also define $N_r = 400$ to be the number of samples for the baseline segment and for a given vector $a$, the notation $a\{n,m\}$ denotes the subvector $\{a_n, \ldots, a_m\}$. Figures 2-3 and 2-4\(^2\) show the locations of the collected recordings for both nets.

![Sensor map](image)

Figure 2-3. Sensor map corresponding to the 128-channel HCGSN setting in Figure 2-1.

### 2.1.1.3 Channel subsets

To simplify the later interpretation of computed results over all possible pairs of recordings, we further propose to group channels into subsets. A standard way of doing so would be to use the international 10-20 system, a system that employs measurements of cranial landmarks for locating electrodes on the scalp [Homan et al., 1987]. Figure 2-5

\(^2\) This figure has been reprinted from page 125 of Electric Geodesics Inc. [2007].
illustrates the 10-20 system and Figure 2-6 groups the electrodes (shown in Figure 2-3) according to a 10-20 scheme.

2.1.2 Signal Processing Overview

A variety of neural activities is recorded at the same time with the ssVEP. This includes muscle activity (EMG), electrocardiac activity (ECG) and electroocular activity (EOG). Added to that come environmental noise from the power supply, the instrumentation and stimulus artefact [Novak et al., 2004]. Hence, most EEG signal processing steps include artefact removal algorithms. Our experimental setting provides an advantage in that the duration of each trial does not exceed 5 seconds. This allowed asking the subject not to blink and hence undermined the effect of EOGs. It also helped keeping muscle activity to a minimum.
Figure 2-5. The 10-20 system is an internationally recognized system to describe the location of EEG electrodes in terms of the distance between the inion and the nasion. F stands for frontal, C for central, P for parietal, T for temporal, O for occipital and A for earlobe reference. © Immrama Institute P.O. Box 16604 Tampa, FL 33687-6604.

Figure 2-6. Sensors in Figure 2-3 grouped according to a 10-20 scheme.

To overcome the effect of line noise, typical notch filters were developed at the noisy components as will be explained in detail in Section 2.2 and Appendix C. On the other
hand, since the presented stimuli have a constant flickering rate, and given the nature of ssVEPs that fluctuate at this specific frequency [Regan, 1989], it is only natural to consider that most relevant information in the recorded signals is included within a frequency band centered at the flickering rate. Such band can be extracted with the use of a bandpass filter. The filter design is controlled by two factors: the order and the quality factor. The latter is defined as the ratio of the center frequency to the bandwidth (Refer to Figure 2-7). Setting these parameters is addressed in the next section. At this stage, it is important to keep in mind that any signal processing step should not distort possible solutions to the problem we are tackling. Again, our goal is to determine whether channel recordings are dependent. Given that any two EEG recordings come with a native delay (because of their spatial location), we need to make sure that any adopted filtering approach does not distort the intrinsic delay information. One way to do so is to use linear-phase filters. Linear-phase filters are desired here because they have constant group delays, i.e. different frequency components have equal delay times. To further elaborate, no distortion is introduced due to the time delay of frequencies relative to one another.

Another point to keep in mind is the ultimate purpose of the experiment at the backend, which is discriminating two cognitive states. It is hence required to analyze the effect of the signal processing approach on the classification performance. Thus we address the bandpassing problem from that perspective and try to find, in Section 2.2.3.3 the set of parameters that maximizes the separability between conditions.

## 2.2 EEG Signal Processing

### 2.2.1 Raw Data

The recorded EEG signals reveal the presence of strong noises at 60-Hz and its odd harmonics (third and fifth). Our filtering approach is two-fold: first, we want to clean the noisy components of the signal and bandpass it to extract the frequency range of interest for dependence analysis and second, we want the chosen set of filtering parameters to minimize the distortion in the signal with respect to measures of dependence. In
Figure 2-7. Standard parameters of a bandpass filter. The quality factor is defined as the ratio of the center frequency to the bandwidth, i.e. $Q = \frac{f_0}{f_2 - f_1}$.

fact, since the filtered output is a convolution of the input EEG with the filter impulse response, one would like to avoid as much as possible ringing that will impact negatively the assessment of channel dependences.

2.2.2 Current Source Densities

Several factors interfere with the process of getting a high quality representation of brain activity when recording an EEG signal. The impact of some of these artifacts can be attenuated like the subject’s head motion and eye blinking, whereas others are uncontrollable such as the electrical activity of some muscles, electrocardiograms and especially the effect of volume conduction. To reduce the effect of the latter, we estimate the strength of extracellular current generators underlying the recorded scalp potentials. This is done using a spherical model of the head based on a linear volume conduction assumption. Such model includes four layers (scalp, skull, cerebro-spinal fluid and brain) and is convenient since it allows derivation of analytic results with low margin of error [Wieringa, 1993]. As a result, we use current source density (CSD) measures that roughly
approximate the locations of the current sources and sinks in a reference-free fashion [Nicholson, 1973; Tenke and Kayser, 2005]. To derive these measures, we follow the procedure outlined by Junghofer et al. [1997] where the calculation of the CSD or the equivalent “Laplacian” method is useful to reduce the spatial lowpass filtering impact caused by the volume conduction property of the tissue, fluids and skull [Codispoti et al., 2006a;b]. This method primarily relies on a sufficient spatial sampling as well as an adequate signal to noise ratio [Junghofer et al., 1997]. These requirements were met for the current study with ERP measurements performed at at least 129 electrode locations with multiple trials per condition. More details about the CSD transform can be found in Appendix A.

2.2.3 Signal Processing

Figure 2-8 shows the recorded signal from one of the channels both in time and frequency domains, corresponding to a face condition. Four clearly discernible spikes can be seen at 17.60, 60, 180 and 300 Hz when analyzing the FFT of the time series consisting of pre- and post-stimulus portions \((N_t^{-1} \sum_{k=1}^{N_t} x^{(72,i)}(1, N_r) \text{ and } N_t^{-1} \sum_{k=1}^{N_t} x^{(72,k)}(N_r + 1, N_s))\). The difference in the estimated \(F_o\) in the graph arises from the discrepancy in the analog display frequency of the monitor.

2.2.3.1 Notch filters

To clean the CSD signal from the line noise artifactual component, we use notch filters. The center frequencies to attenuate are at \(f_1 = 60\) Hz, \(f_2 = 180\) Hz and \(f_3 = 300\) Hz. The quality factor, defined for a filter as \(Q = f_c/\Delta f\) (where \(f_c\) denotes the filter’s center frequency and \(\Delta f\) the width of the frequency band at 3 dB) is chosen to be \(Q_1 = 20\) for the first notch filter. Moreover, since we wish the three filters to have the same 3-dB bandwidth \(\Delta f\), we set \(Q_2 = 3 \times Q_1 = 60\) and \(Q_3 = 5 \times Q_1 = 100\). We use Butterworth filters because they have a more linear phase response in the passband than a Chebyshev Type I or Type II filter or an elliptic filter and show less ringing (Refer to Appendix B).
2.2.3.2 Bandpassing

When computing the dependence between channel locations, the frequency band around $F_o$ is of particular relevance to us since it carries information about the response to the visual stimuli at each recording site; hence a suitable bandpassing scheme is needed to extract the flickering frequency band. Since we desire constant group and phase delays for the estimation of dependence, a linear-phase filter is appropriate. Moreover, as we propose to study the effect of different filtering schemes with respect to the dependence backend, we proceed using a simple non-interactive method that is optimal with respect to the square error criterion, namely the linear-phase least-square FIR filter. One of the advantages of such approach is that it is suitable for arbitrary desired amplitude response. For such filter, the problem can be stated as follows:

$$h_{N,Q} = \min_h \| \Gamma h - D \|^2$$

(2-1)
where $h_{N,Q} = [h(n)]$ denotes the impulse response of a Type II (even symmetry) filter of length $N+1$ ($N$ even), $D = [D(w_r)]$ is a length $R$ vector containing the ideal response at a set of frequencies $\{w_r\}$. $D$ implicitly defines the quality factor $Q$ and the center frequency $f_c$. Row $p$ of $\Gamma$ is given by:

$$ \Gamma^p = [1, 2\cos(w_p), \ldots, 2\cos(w_p(N-1))] \text{ where } p \in [0, \ldots, N-1] \quad (2-2) $$

The expression of $\Gamma$ in Equation 2-2 can be obtained by expressing the filter as:

$$ H(w_r) = h_{N,Q}(0) + 2\sum_{n=0}^{N/2} h_{N,Q}(n)\cos(w_rn) \text{ with } r \in [0, \ldots, N-1] \quad (2-3) $$

where the above step follows from the even symmetry property and shifting by $N/2$ samples. The problem reduces to a least-square optimization that can be formulated by:

$$ \hat{h}_{N,Q} = \arg\min_{h_{N,Q}} \|\Gamma h_{N,Q} - D\| \quad (2-4) $$

Hence:

$$ \hat{h}_{N,Q} = [(\Gamma^T\Gamma)^{-1}\Gamma^T]D \quad (2-5) $$

The obtained estimate is used to generate bandpass filters having different orders and quality factors.

2.2.3.3 Testing

It is easy to observe that decreasing the filter’s bandwidth, or alternatively increasing its quality factor, increases the dependence between two channels of filtered signals, since
both these signals become close to a single frequency sinusoid with a specific frequency i.e. the flickering frequency. ³

On the other hand, filters with low quality factors are more affected by noise because of the larger bandwidth, and as a result, dependence values between the filtered channels drop. Therefore, to ensure robustness it should be established that there exists a sufficiently large set of parameter values where the dependences captured among the filtered signals are meaningful and stable. Moreover since our goal is to maximize the separability between the two conditions, face and Gabor patch, we consider it to be a criteria to judge the robustness of the effect of filter variation.

We hence compute pairwise dependence values per condition, per trial and per filter, and then apply statistical tests to see how effectively we can differentiate the two conditions for each filter. The statistical test used for this purpose is the two-sample Kolmogorov-Smirnov (KS) test, which is a non-parametric test to compare two sample vectors. The KS test tries to estimate the distance between the empirical distribution functions of the two sets of samples. It is described in more details in Chapter 6. The complete testing procedure is detailed in Algorithm 1 for a measure of dependence that will be addressed in more details in Chapter 3 (Section 3.2): the generalized measure of association (GMA).

**Impact of filter parameters and backend:** Following the procedure described in Algorithm 1, we obtain the plot shown in Figure 2-9. We present the results for only 3 different filter orders out of 20 between 10 and 300, since the effect of the filter order on the shape of the curve is minimal. We have evaluated the curve for 250 quality factors in the range ~0.1 and 175. Since the readings of any two pair of channels is affected by the delay in signal propagation, we embed the signal in $m = 8$ dimension before computing the

³ Besides, filters with higher quality factors induce better stopband attenuation but exhibit more ringing.
Algorithm 1: Testing Procedure with Different Filters

**Input:** \( O \) is a vector of \( M \) filter orders, \( f \) is a vector of \( N \) filter bandwidths, \( \gamma \) is the embedding dimension

**Output:** An \( M \times N \) matrix \( R \) of pairwise KS test decisions per filter order per quality factor

1 begin
2 \hspace{1em} for \( m \in \{1, \ldots, M\} \) do
3 \hspace{2em} for \( n \in \{1, \ldots, N\} \) do
4 \hspace{3em} Compute \( h_{O(m),f(n)} \) as in (2–5).
5 \hspace{3em} for \( i \in \{1 \ldots N_c\} \) do
6 \hspace{4em} for \( j \in \{1 \ldots N_c\} \) do
7 \hspace{5em} for cond \( \in \{\text{Gabor, Face}\} \) do
8 \hspace{6em} for \( k \in \{1 \ldots T\} \) do
9 \hspace{7em} Compute \( Y^{(k)\text{(cond)}}_{m,n,i,\cdot} = h_{O(m),f(n)} \star X^{(k)\text{(cond)}}_{i,\cdot} \) and
10 \hspace{7em} \( Y^{(k)\text{(cond)}}_{m,n,j,\cdot} = h_{O(m),f(n)} \star X^{(k)\text{(cond)}}_{j,\cdot} \), where \( \star \) denotes convolution.
11 \hspace{7em} - Embed filtered series \( Y^{(k)\text{(cond)}}_{m,n,i,\cdot} \) and \( Y^{(k)\text{(cond)}}_{m,n,j,\cdot} \) in \( \gamma \) dimension to get \( \xi_1 \) and \( \xi_2 \)
12 \hspace{7em} - Set \( D^{(k)\text{(cond)}}_{m,n,i,j} = \text{GMA}(\xi_1, \xi_2) \)
13 end
14 end
15 Compute \( R_{m,n} \) as \( \sum_i \sum_j \text{KS}(D^{\cdot,\text{Face}}_{m,n,i,j}, D^{\cdot,\text{Gabor}}_{m,n,i,j}) \).
16 end
17 end
18 end
19 end
20 end
21 end
22 end

GMA values. \( m = 8 \) corresponds to 32 ms. More details are included in [Fadlallah et al., 2011].

The curve starts with relatively low values corresponding to quality factors in the range \( Q < 0.4 \), then stays almost stable in the range \( Q \in [0.6, 1.5] \), and then decreases again. This is inline with our expectations since for a low quality factor, the dependence level is reduced due to the presence of noise and the method looses discriminability, whereas on the other hand, for a high quality factor the dependence values are usually high, and therefore the method again perform worse. The stable performance of the
Figure 2-9. Plot showing the change in the number of channels for which the two conditions are correctly discriminated by the KS test as the bandpass filter quality factor increases. Tests were performed for 3 different filter orders and GMA was used as dependence measure. The curve increases with the quality factor, reaches a plateau and then decreases as the quality factor goes up, clearly showing a region where the estimated dependence values are robust.

method in the range $Q \in [0.6, 1.5]$ can be justified since for this range of quality factors the bandwidth range becomes $[\sim 11, \sim 29]$. It is possible that this particular bandwidth range covers the modulation of the flickering frequency well, thus extracting the essential information for discriminating the stimuli.

Before proceeding further, we check the dependence values for two pairs of channels that we expect to be highly and less dependent, respectively. Figure 2-10 shows the GMA values between a pair of channels in the occipital area (near $O_2$), and a pair having one channel in the occipital (near $O_2$) and the other in the frontal (near $F_{p1}$) area. As expected, we observe that GMA increases with the quality factor for the channels likely to be dependent since the effect of noise fades with the reduction of the passband. For the channels less likely to be dependent, the reduction of the noise level is masked by the differences between the two signals in the passband.
Figure 2-10. GMA vs. quality factor plot for 3 different filter orders for two channels having (a) high dependence and (b) low dependence.

Results: The obtained ssVEP signal after filtering is shown in Figure 2-11. A clear difference in frequency spectra can be noticed between the baseline and post-stimulus portions. Inline with our expectation of higher power at $F_o$ after stimulus inception, the first FFT magnitude plot shows peaks in the higher alpha-range whereas the second has a stronger and more marked peak at the flickering frequency.

Figure 2-11. Filtered CSD signal at channel location 72 averaged over the 15 trials. (a) Signal in time prior and after stimulus onset. (b) Frequency components of the baseline segment. (c) Frequency components of the post-stimulus segment.
2.3 Analysis in the Frequency-Domain

We first analyze the filtered signals in the frequency domain over the sensor space to assess the filtering impact and characterize the participants activity in the $F_o$ frequency range. The analysis is restricted to the frequency domain and is based on all recorded time samples, thus presenting limitations in time resolution.

2.3.1 Power

Power is visualized for different frequency bands between 16 and 18.9 Hz. For both conditions, the power in a 0.5 Hz band including $F_o$ clearly dominates those in the other ranges and is mostly concentrated in the right occipito-temporal regions. This is shown in Figure 2-12 where power was averaged over all trials and plotted by sensor location.

![FFT power in electrode space for filtered signals averaged over trials. Upper row corresponds to the face condition and lower row corresponds to the Gabor patch condition. Both conditions show distinctively high power in the flickering frequency band mostly concentrated in the right occipito-temporal region.](image)

The variability per trials of the FFT power at the flickering frequency is shown in Figure 2-13. The computed variance per trial is inline with the power distribution plot (Figure 2-12) that shows more localization of the power at 17.5 Hz for the face condition. Figure 2-13 (b) illustrates two different linear fits between the variance vectors quantiles with standard normal quantiles for the two conditions.
2.3.2 Phase

Looking at the power of the FFT signal is not enough to infer information about the dynamics of the signal in sensor space at $F_o$ and thus describe the temporal characteristics of the data. A knowledge of the phase is necessary to extract such information. For the FFT $S(f)$ of a time series $s(n)$, the phase is defined as $\phi(f) = \tan^{-1}[\Im\{S(f)\}/\Re\{S(f)\}]$ and yields the phase relative to the start of the time-domain signal. Translating phase differences between the FFTs of two signals $S_1(f)$ and $S_2(f)$ at a frequency $f_x$ into a time delay can be done using $\delta t_{s_1-s_2}(f_x) = [\phi_{S_1}(f_x) - \phi_{S_2}(f_x)]/[2\pi f_x]$. We should however be cautious when using such conversion unless we have access to the true unwrapped phase, which is difficult when the studied signal suffers from background or intrinsic noise [Lachaux et al., 1999; Varela et al., 2001]. In Figure 2-15, we plot the FFT phase per location and condition for the $[17.2 - 17.5]$ Hz range, where a well-defined narrowband activity exists and neglect the other bands since they display no power spectrum peaks.
Figure 2-14. Power distributions for the two conditions. Error bars represent variability with trials. The $\text{KS}$ statistic obtained is 0.2891 with a $p$-value of 0.000316. The $\text{KS}$ statistic is discussed in Chapter 6.

Figure 2-15. FFT phase in electrode space for filtered signals averaged over trials. Left map corresponds to the face condition and right map to the Gabor patch condition. Regions with lighter colors correspond to lower phase values. Phase wrapping makes it difficult to draw a correct interpretation in terms of time delays between different channels.

The phase plots do not provide any info in the range of interest. Even if we assume otherwise, the time resolution provided via this approach is very limited. Thus, in order to quantify the dependence between electrode locations in response to the visual stimuli, we apply dependence measures on time windows of the filtered signals. Such approach
bypasses the need to extract time delays from the signal’s phase. Besides, it performs better than a classifier based only on power discrimination (as illustrated in Figure 2-14). We hence tackle the problem from a time-domain perspective, using directly the time series processed by the linear-phase FIR filter previously discussed to provide a better discrimination between the two conditions. Another reason to motivate this approach is the difficulty to infer patterns that differentiate the two conditions (Face and Gabor) from FFT descriptors as shown in Figure 2-16, for all subjects.

![Graphs showing average power at 17.5 ± 0.3 Hz for each channel across the six subjects. The time series were averaged over trials prior to computing the FFT power. 129 EEG recordings have been collected from subjects 1, 2 and 3 and 257 recordings have been collected from subjects 4, 5 and 6. It is difficult to infer patterns that differentiate the two conditions (Face and Gabor) from FFT descriptors.]

Figure 2-16.
CHAPTER 3
MEASURES OF DEPENDENCE

In this chapter, we describe measures of dependence that can be applied in the general context of neural data and the specific context of this dissertation. A function approach to measures of dependence has been proposed by Renyi [1959], who proposed seven postulates to define a measure of dependence between two random variables. We start by outlining traditional measures of dependence such as correlation, mutual information, phase synchrony and standard measures of association, before discussing measures of causality, a family of measures that provides directionality information besides dependence. We then move to discuss two novel measures of dependence: the first is based on the notion of generalized association and the second on measures of signal complexity.

3.1 Measures of Dependence in the Literature

3.1.1 Dependence

Dependence can be simply interpreted as the absence of independence. Many applications in engineering, statistics, econometrics, geophysics, biology and medicine require, whenever two random variables are not independent, to determine the degree of dependence between the two random variables. In this section, we give a brief sketch of the most used measures of dependence. Three correlation measures and two information theoretic measures are presented, in addition to Granger causality, phase synchrony, and coherence-based measures. Note that the latter can be categorized as frequency-based measures but are still reported here because of their relatively popular use in neural contexts.

3.1.1.1 Correlation

Correlation tries to measure the relation between two or more variables and is probably the most widespread used measure of dependence. The common understanding of correlation refers to linear correlation, that can be estimated using Pearson’s correlation coefficient. Other measures of correlation however transcend this restriction.
**Pearson's r:** Pearson’s correlation [Pearson, 1896] is obtained for two variables $X = \{x_i\}_{i=1}^{n}$ and $Y = \{y_i\}_{i=1}^{n}$ by dividing the covariance of the two variables by the product of their standard deviations. It can be defined as:

$$C_{XY} = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$  \hspace{1cm} (3-1)

where in the above equation, the covariance between $X$ and $Y$ is defined as $\text{cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$. Pearson’s correlation ranges between -1 and +1. The first reflects a perfect negative linear relationship whereas the second describes a perfect positive linear relationship. A correlation of 0 is obtained if $X$ and $Y$ are statistically independent, but can also be obtained if $X$ and $Y$ are uncorrelated but not independent. Replacing the expected value operator by the sample mean, we obtain the sample correlation coefficient that can be defined as:

$$r_{XY} = \frac{n \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2}}$$  \hspace{1cm} (3-2)

**Spearman’s rho:** Spearman rank correlation (or Spearman’s rho) is a non parametric measure of the degree of association between two variables $X$ and $Y$. It can be obtained by applying Pearson’s correlation coefficient on ranks of the input variables. More details about the power of Spearman’s rho have been discussed by Siegel and Castellan [1988]. For the two variables $X$ and $Y$, it can be defined as:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$  \hspace{1cm} (3-3)

where $d_i$ represents the difference between the ranks of the two observations $x_i$ and $y_i$:

$$d_i = \text{rank}\{x_i\} - \text{rank}\{y_i\}$$  \hspace{1cm} (3-4)
**Kendall’s tau:** Kendall’s rank correlation (or Kendall’s tau) addresses some of Spearman’s rank correlation 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 occurrence probability of concordant and discordant pairs. It is generally referred to as Kendall’s tau and can be defined as:

\[
\tau_{XY} = \frac{N_c - N_d}{\frac{1}{2}n(n-1)}
\]

where \(n\) denotes the time series length and \(N_c, N_d\) refer respectively to the number of concordant and discordant pairs, in turn defined for two variables \(X\) and \(Y\) as:

1. \(N_c\) corresponds to either of the following cases where:
   - \(\{x_i > x_j \text{ and } y_i > y_j\}\)
   - \(\{x_i < x_j \text{ and } y_i < y_j\}\)

2. \(N_d\) corresponds to all other cases where:
   - \(\{x_i > x_j \text{ and } y_i < y_j\}\)
   - \(\{x_i < x_j \text{ and } y_i > y_j\}\)

**Discussion:** Spearman’s rho and Kendall’s tau can capture monotone dependence rather than just linear dependence [Nelsen, 2002]. The following inequality captures the relationship between the two [Siegel and Castellan, 1988]:

\[-1 \leq 3 \times \text{Kendal’s } \tau - 2 \times \text{Spearman’s } \rho \leq 1\]  

(3–6)

On the other hand, Kendall’s tau and Spearman’s rho mean two different things. The latter can be thought of as the regular Pearson product-moment correlation coefficient applied on the ranks, whereas Kendall’s tau represents a probability. It is specifically the
difference between the probabilities that the observed data share the same order or have different orders [Hill and Lewicki, 2007].

### 3.1.1.2 Mutual information

Mutual information has been championed by many researchers as a measure of dependence, mainly because of the influential work of Shannon [1948] in proposing mathematical interpretations for concepts of uncertainty and information.

**Definition:** Assuming that $X$ and $Y$ have a joint probability distribution $P_{XY}(x,y)$, the mutual information between the two variables is defined as:

$$ I(X,Y) = \sum_{x,y} P_{XY}(x,y) \log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)} $$

$$ = \mathbb{E}_{P_{XY}} \left[ \log \frac{P_{XY}}{P_XP_Y} \right] $$

where $\mathbb{E}_X$ represents the expected value operator over distribution $X$ and $P_X(x)$ and $P_Y(y)$ denote the marginal distributions: $P_X(x) = \sum_y P_{XY}(x,y)$ and $P_y(y) = \sum_x P_{XY}(x,y)$. An alternative definition of mutual information using the definition of the entropy and conditional entropy quantities is:

$$ I(X,Y) = H(X) - H(X|Y) $$

Mutual information can be then seen as the reduction in a variable’s uncertainty when the second variable is known. In Equation 3–8, the marginal entropy $H(X)$ (resp. $H(Y)$) and conditional entropy $H(Y|X)$ (resp. $H(X|Y)$) can be respectively expressed as:

$$ H(X) = \sum_{x \in X} P_X(x) \log P_X(x) $$

and:

$$ H(Y|X) = \sum_{x \in X, y \in Y} P_{XY}(x,y) \log \frac{P_X(x)}{P_{XY}(x,y)} $$

(3–10)
**Estimation:** Estimating mutual information is a challenging problem. In this dissertation, we use the estimator proposed by Kraskov et al. [2004]. Unlike conventional mutual information estimators based on binning or kernel density estimation (KDE), this method relies on entropy estimates using the $k$-nearest neighbour (kNN) algorithm. The expression of the estimator is shown below:

$$\hat{I}(X,Y) = \psi(k)\phi[\psi(n_x + 1) + \psi(n_y + 1)] + \psi(n) \quad (3-11)$$

where $\phi[] = \langle \cdot \rangle = N^{-1} \sum_{i=1}^{N} E[\cdot(i)]$, $\psi(x)$ is the digamma function and $n_c(i)$ for a random variable $E$ represents the number of points $e_j$ whose distance from $e_i$ is strictly less than a certain distance $\epsilon_i$, in turn computed as the difference between $z_i$ ($Z$ denoting the maximum norm over the space consisting of the two random variables) and its $k^{th}$ neighbour. In this dissertation, we use a value of $k = 4$, after several trials. Another way of writing this estimate of mutual information is shown in Equation 3–12.

$$\hat{I}(X,Y) = \psi(k) + \psi(N) - \langle \psi(n_x) + \psi(n_y) \rangle \quad (3-12)$$

### 3.1.2 Measures of Causality

Causal interpretations of random variables have recently become an area of increased focus [Edwards, 2000; Patterson, 2006]. However most common applications using causality lie within the context of time series, and this will be the assumption along the lines of the following definitions.

#### 3.1.2.1 What is causality?

Defining causality has been a controversial issue for centuries. Even today, there is still no clear answer to how causality is defined. Narrowing down the span of the question, we might find definitions of causality that satisfy (or not) the interests of scientists in that area.
In Signal Processing, a causal System is a system where the output depends on past/current inputs but not future inputs. i.e. the output $y(t_o)$ only depends on the input $x(t)$ for values of $t \leq t_o$. Similarly, a discrete-time system is said to be causal if, for any given $m$, the output sequence at the index $n = m$ depends only on the input sequence for $n \leq m$. Causality is a mandatory factor in the design and implementation of real-time DSP systems for an obvious reason: all real-time systems must be causal as they do not have future inputs available to them [Patterson, 2006]. In this manuscript, we denote causality by $C$, i.e. $X C Y$ means $X$ causes $Y$, or $X$ and $Y$ are related through a causal system. Causality is transitive in the sense that if $E C F$ and $F C G$ are true, then $E C G$ is true as well. In other terms, feeding the output of causal System $A$ to the input of causal System $B$ means that the input of $A$ causes the output of $B$, or alternatively that the combined system $A - B$ is causal.

### 3.1.2.2 Cross-spectral analysis

It is crucial for many applications to know how two time series are related. Cross-spectral methods essentially represent an extension of spectral analysis to time-series, trying to determine the correlations between two time-series. They are particularly useful to describe relationships among variables when one is causing the others. It has been used for several applications including communications engineering, econometrics, geology and climatology.

The cross-spectrum function is none other than the Fourier transform of the cross-covariance function. It is hence a complex valued function that provides information about the distribution of two processes activity across their frequency components. In the following, let $\gamma_{xy}(\tau)$ and $\Gamma_{xy}(f)$ denote respectively the cross-covariance function of two time series $X$ and $Y$ and their cross-spectrum function.

$$
\Gamma_{xy}(f) = \mathcal{F}\{\gamma_{xy}\}(f) = \sum_{\tau=-\infty}^{\infty} \gamma_{xy}(\tau) e^{-2\pi i \tau f}
$$

(3-13)
**Coherence:** With the above definition of the cross-spectrum, we can define the coherence function in an analogue way to the conventional correlation coefficient, i.e. as the ratio of the squared magnitude of the cross-spectral density and the product of the auto-spectral densities. This function can be seen as an indicator of how good the contribution of \( x(t) \) is in predicting \( y(t) \) using an optimum linear least squares function. The advantage coherence has over correlation consists in being a function of frequency, which allows to track correlations across frequency bands. Coherence achieves values between 0 and 1 and works best under stationarity assumption.

\[
C_{xy} = \frac{|\Gamma_{xy}|^2}{\Gamma_{xx}\Gamma_{yy}}
\]  \hspace{1cm} (3-14)

From the cross-spectrum, we can also define a phase information indicator \( \phi_{xy}(f) \) at each frequency \( f \) as:

\[
\phi_{xy}(f) = \tan^{-1} \frac{\Im \{\Gamma_{xy}(f)\}}{\Re \{\Gamma_{xy}(f)\}}
\]  \hspace{1cm} (3-15)

The estimate returned by the relative phase information should only be considered when the coherence magnitude is significant. For two correlated time series (or processes), \( \phi_{xy}(f) \) will show a preferred relative phase angle as function of frequency [Bressler, 2008].

**Coherence-based methods:** Saito and Harashima [1981] proposed a new measure of coherence they referred to as directed coherence (DC) to determine the information flow direction and directed spectral characteristic of EEG signals in a two-channel model. The measure was introduced by decomposing the standard coherence function into two components, that can be labeled directed coherences, representing the feedforward/feedback aspects of the interactions between the time series.

A more accurate extension of directed coherence within the context of autoregressive (AR) models resulted in directed transfer function (DTF), introduced by Kaminski and
Blinowska [1991]. In 2001, Baccala and Sameshima [2001] introduced the concept of partial directed coherence (PDC). PDC provides a frequency-domain representation of multivariate processes. PDC is based on the concepts of Granger causality (to be discussed shortly), DC and DTF.

Applications: Spectral methods have been used by Scheeringa et al. [2011] to analyze how the blood-oxygen-level-dependent (BOLD) signal in humans is related to neuronal synchronization across different frequency bands when performing a specific cognitive task, by Thatcher et al. [1986] to explain differences in connectivity between different brain regions by computing EEG coherence as a function of different interelectrode distances, and by Tucker et al. [1986] to measure the degree of functional connectedness among right hemisphere regions.

Applications of DC on EEG data have been proposed by Wang and Takigawa [1992], Wang et al. [1992], Inouye et al. [1995] and Korzeniewska [2003]. Contributions have also been made using DTF [Astolfi et al., 2004; Babiloni et al., 2005; Kaminski et al., 2001] and PDC [Astolfi et al., 2006; Schelter et al., 2006].

3.1.2.3 Granger causality

Many real-world situations present a two-way causality. In such cases where feedback occurs, using coherence and phase diagrams might not be useful. Granger causality [Granger, 1969] addresses this shortcoming of early spectral methods and has been perhaps the most widely-established means of identifying causal relationships between times series [Hesse et al., 2003].

Intuition: The idea behind Granger Causality can be understood from the following illustrations:

1. Given: $X$, $Y$ and $W$. If the forecast of $X_{t+1}$ given $X_t$, $Y_t$ and $W_t$ is found to be more accurate than the forecast of $X_{t+1}$ given $X_t$ and $W_t$, we deduce that $Y_t$ has information useful in forecasting $X_t + 1$ that cannot be deduced from $X_t$ and $W_t$. In this case, we say that $Y_t$ Granger-causes $X_{t+1}$ under the below conditions:
(a) $Y_t$ occurs before $X_{t+1}$

(b) $Y_t$ contains information useful in forecasting $X_{t+1}$ not available from other variables.

2. A variable $X$ Granger-Causes $Y$ if $Y$ can be better predicted using the histories of both $X$ and $Y$ than it can using the history of $Y$ alone.

3. $X$ does not Granger-Cause $Y$ ($X \not\text{GC} Y$) if and only if the prediction of $Y$ based on the universe $U$ of predictors is not better than the prediction of $Y$ based on the set of predictors $U - X$, i.e. on the universe with $X$ omitted.

**Formulation:** Mathematically, let our information set be denoted by $I_t = \{X_t, Y_t, Z_t, W_t\} = \{x_t, y_t, z_t, w_t; x_{t-1}, y_{t-1}, z_{t-1}, w_{t-1}; \ldots; x_1, y_1, z_1, w_1\}$. Other variables can be contained in $I_t$ as well.

**Definition.** $y_t$ is said not to be Granger Caused by $x_t$ with respect to the information set $I_t$ if:

$$\mathbb{E}[y_{t+h}|I_t] = \mathbb{E}[y_{t+h}|I_t - X_t] \quad \forall \; h.$$  \hfill (3–16)

Equivalently, $X_t \text{ GC} Y_t$ if the linear predictor of $y_{t+h}$ based on $I_t$ has different expected value than the optimal linear predictor of $y_{t+h}$ based on $y_t, y_{t-1}, \ldots, y_1$ for any $h$.

**Shortcomings:** The main shortcoming of Granger causality is that it can only give information about linear features of signals. Hence it quantifies *linear causality in the mean*. Also, Granger causality does not imply true causality. To illustrate this point, consider two processes driven by a common third at different lags. In such case, the alternative hypothesis might still be not rejected.

**Applications:** Seth [2010] recommends the following measures to be performed prior to applying Granger Causality on EEG data:

1. Artifact reduction.
2. Downsampling.
3. Re-referencing to a common source.
4. Subtraction of pre-stimulus baseline.
Other measures of causality based on information theoretic concepts have also been proposed. Directed information [Quinn et al., 2011] has been suggested for example as an information theoretic quantity to infer robust measures of statistical causality from spike train populations.

3.1.3 Measures of Synchrony

Although no consensus exists on the definition of synchronization, it is rather acceptable to formulate phase synchrony as the degree of phase locking between oscillators. To illustrate the ideal case, let us assume we have two oscillators with phases $\phi_1$ and $\phi_2$ at frequency $f$. $\phi_1$ and $\phi_2$ usually vary with time but to further simplify, we consider a case where they are constant. We say that the oscillators are synchronized at $f$ if it is possible to find two integers $n$ and $m$ that verify the relation $n\phi_1 - m\phi_2 = \text{const}$ [Rosenblum et al., 2010]. Phase synchronization of chaotic oscillators can be witnessed in several complex systems such as the human brain, especially when it performs cognitive processes. As a result, measures of synchrony have been widely used in a variety of disciplines including neural science and cognitive psychology.

3.1.3.1 Phase-locking statistics

Quantifying frequency-specific synchronization between two neuroelectric recordings has been first suggested by Lachaux et al. [1999]. Assume we have two time series $x(t)$ and $y(t)$, the first step towards computing the phase synchrony between the two series is extracting their instantaneous phases $\phi_x(t)$ and $\phi_y(t)$. Two main methods are used for that purpose: the first uses the analytic signal and the Hilbert transform:

$$\phi_x(t) = \arctan \left(\frac{\tilde{x}(t)}{x(t)}\right)$$  \hspace{1cm} (3–17)

where:

$$\tilde{x}(t) = \frac{1}{\pi} \text{P.V.} \int_{t'-\infty}^{t'+\infty} \frac{x(t')}{t-t'} dt'$$  \hspace{1cm} (3–18)
In 3–18, P.V. refers to the Cauchy principal value or principal value integral. Such representation of the phase is possible since the analytic signal can be written as:

\[ x_a(t) = x(t) + i\tilde{x}(t) = A_x e^{i\phi_x(t)} \]  

(3–19)

Wavelet-based techniques constitute another class of phase synchrony estimation methods. Lachaux et al. [1999] suggest to convolute the signal first with a Gabor wavelet \( G(t, f) \), centered at frequency \( f \), which is defined at time \( t \) as:

\[ G(t, f) = \exp \left( \frac{t^2}{2\sigma_t^2} \right) \exp(j2\pi ft) \]  

(3–20)

where \( \sigma_t \) is defined as a constant multiple of \( 1/f \). We then compute the quantity \( \theta(t, n) \) to define the phase difference between recording sites \( i \) and \( j \):

\[ \theta(t, n) = \phi_i(t, n) - \phi_j(t, n) \]  

(3–21)

where \( n \) refers to the trial index, and \( \phi(t, n) = \arg(x(t) * G(t, f)) \) is the phase of the signal’s convolution with the wavelet. The phase locking value (P.L.V.) can be then defined as:

\[ \text{P.L.V.}(t) = \frac{1}{N} \left| \sum_{n=1}^{N} \exp(j\theta(t, n)) \right| \]  

(3–22)

A similar approach based on the Morlet wavelet was discussed by Gruber et al. [2001].

### 3.1.3.2 Mean phase coherence

Another measure of phase synchronization based on the distance between phase values is proposed by Mormann et al. [2000]. To compute the mean phase coherence, the instantaneous phase is first extracted from a pair of signals using the Hilbert transform.
The estimated values are then projected to a unit circle and the mean phase coherence is returned as a functional of the distances between the instantaneous phases.

\[
\text{M.P.C.} = \left| \frac{1}{N} \sum_{k=1}^{N-1} \exp^{j(\phi_x(k/F_s) - \phi_y(k/F_s))} \right|
\]  

(3–23)

where \( F_s \) denotes the sampling frequency and \( N \) the number of samples in the time series.

An alternative way of writing the mean phase coherence is:

\[
\text{M.P.C.} = \sqrt{\mathbb{E} [\cos(\phi_x(t) - \phi_y(t))]^2 + \mathbb{E} [\sin(\phi_x(t) - \phi_y(t))]^2}
\]  

(3–24)

where \( \mathbb{E}[.] \) denotes the expectation operator.

### 3.2 Novel Measures of Dependence

In the second part of this chapter, we discuss two recently proposed measures that can be used to infer dependence between two time series. The first starts by generalizing the concept of association for time series and the second is based on symbolic dynamics.

Before discussing these measures, we introduce some preliminary concepts.

#### 3.2.1 Preliminary Concepts

##### 3.2.1.1 Dependence

Renyi [1959] proposed seven axioms a proper measure of dependence \( Q(X, Y) \) should satisfy, where \( X \) and \( Y \) represent two random variables [Príncipe, 2010]. According to Renyi, if \( Q(X, Y) \) is a measure of dependence, then \( Q(X, Y) \) is:

1. Well defined over \( X \) and \( Y \).
2. Symmetric or \( Q(X, Y) = Q(Y, X) \).
3. Bounded or \( 0 \leq |Q(X, Y)| \leq 1 \).

\(^1\) A measure of dependence can be allowed to be asymmetric to infer causal relationships i.e. effective connectivities among data.
4. Zero for independent random variables, i.e. \( Q(X,Y) = 0 \) iff \( X \) & \( Y \) are independent.

5. One if a random variable is a function of the other, i.e. \( Q(X,Y) = 1 \) if \( Y = f(X) \).

6. Invariant or \( Q(X,Y) = Q(f(X), g(Y)) \) for any functions \( f, g \).

7. \( Q(X,Y) = \rho(X,Y) \) for \( X,Y \sim N(0,\sigma^2) \), where \( \rho \) denotes the correlation coefficient.

Unlike other approaches to define dependence, these postulates lead to the notion of “functional” dependence.

### 3.2.1.2 Association

There are two mainstream methods to estimate dependence between two random variables. The first starts by constructing a function of the random variables under certain criteria then tries to build a consistent estimator for that measure [Kruskal, 1958]. The second starts from the samples, finds a way to capture attributes from these samples, then formulates a measure for which the proposed method is an estimator [Seth, 2011]. Working with samples circumvents the difficulty to find a suitable and consistent estimator when starting with the random variables themselves. The tradeoff here is between two things. The first is to have a clear and well understood concept of dependence for which it is difficult to find an estimator satisfying the different properties. The second is to start from the realizations and define the measure in a build-up fashion. In most practical situations, we only have access to a finite number of realizations, and because of this limitation, understanding the properties of an estimator can be seen as more important than the measure. The concept of association provides a clear understanding of dependence both in the context of random variables and realizations since it treats realizations as a probability law by assigning equal masses over the realization values [Whitt, 1976]. However the concept of association is only restricted to random variables assuming values on the real line. Extending the concept of association from \( \mathcal{R} \) to metric spaces has been proposed by Seth [2011] since such spaces are often encountered in practical problems.
Motivation: As discussed in the beginning of this chapter, both Spearman [1904] and Kendall [1938] generalized the concept of correlation to capture monotone instead of just linear dependence. To satisfy the desired properties of an estimator of dependence, the concept of association is extended from $\mathcal{R}$ to metric spaces. Inspired by the idea of association, and given realizations $(\{x_i, y_i\}_{i=1}^n)$ from a pair of random variables $(X, Y)$, $Y$ can be said to be associated to $X$ if close realization pairs of $Y$ (or $\{y_i, y_j\}$) are associated with close realization pairs of $X$ (or $\{x_i, x_j\}$). Closeness in this context is defined in terms of the distance metrics of the spaces where the realizations lie in ($X$ and $Y$). Such approach to dependence is particularly appealing since it is intuitive, and does not require the domain of the random variables to have an order. In addition, it allows realizations to exist in more abstract domains such as vectors or point processes [Seth et al., 2010], and therefore, defines dependence irrespective of the nature of these domains.

The corresponding estimator is referred to by the term generalized measure of association (or GMA for short). GMA generalizes the concept of association by considering the distance between realizations rather than their absolute locations, retains certain desired invariance properties of a measure of association, and only requires that they be defined in a metric space. It has the advantage of being a parameter-free method. From a time complexity perspective, it runs in $O(n^2)$ in the worst case scenario. In this work, GMA is applied on real valued sequences. A detailed formulation of GMA is provided in the next section.

Formulation: Now that the concept is roughly determined, the next step would be to design an estimator that satisfies the desired properties. This can be done by letting the estimator only consider the relative positions of the realizations with respect to each other rather than their absolute locations. Again, this is inspired by the approach considered by Kendall or Spearman compared to the approach considered by Pearson. We
first consider a simple illustration of that by performing the following two steps for all
indexes $i \in \{1, \ldots, n\}$:

1. Find $x^*_j$ closest to $x_i$ in terms of $\delta_X$.
2. Find the rank $r_i$ of $y^*_j$ in terms of $\delta_Y$

The ranks $r_i$ in the second step can be obtained via $r_i = \# \{ j : j \neq i, \delta_Y(y_j, y_i) \leq \delta_Y(y^*_j, y_i) \}$. The notation $\delta_Z$ denotes the associated metric of the space $Z$. The next step would be to consider the computed $r_i$’s as realizations of a random variable we refer to as the rank variable $R$. Under independence of $X$ and $Y$, $R$ should have a uniform distribution. On the other extreme, if the two variables have a perfect linear relationship, the PDF of $R$ will assign a probability of 1 when $r_i = 1$. In an intermediate scenario, the $r_i$’s would be skewed towards a lower value towards 1. Hence we can say that the more the skewness of the distribution of $R$, the more random variables $X$ and $Y$ are dependent. The final step in dependence estimation is capturing the skewness of $R$ by simply computing the area under its CDF and normalizing by $(n - 1)$. This measure is called generalized measure of association or GMA [Seth, 2011] and the values it estimates range from near 0.5 for independent random variables and 1 for dependent ones. The set of points $x^*_j$ closest to $x_i$ in terms of the associated metric $\delta_x$ can be defined as follows:

$$j^* = \arg \min_{j \neq i} \delta_x(x_i, x_j) \quad (3-25)$$

Algorithm 2 describes in detail the steps involved in computing GMA between two time series. In Algorithm 2, the spread of the ranks is computed to address cases where two or more realizations share the same distance from a third one. Upon completion of the algorithm, $\sum P(R = r) = 1$ and hence corresponds to a valid PDF. An alternative way of formulating GMA would be:

$$GMA = \frac{1}{n - 1} \sum_{r=1}^{n-1} (n - r)P(R = r), \quad (3-26)$$
Algorithm 2: Generalized Measure of Association (GMA)

Input: Realizations pairs \( \{x_i, y_i\}_{i=1}^{n} \) assuming values in the joint space \( X \times Y \)

Output: Estimated dependence \( d \in [0.5 : 1] \)

begin

Initialization:

Assign \( P(R = r) = 0 \ \forall \ r \in \{1, \ldots, (n-1)\} \)

for \( i \in \{1 \ldots n\} \) do

- Find \( x_{j^*} (j^* \in J) \) closest to \( x_i \), equivalently \( j^* = \arg \min_{j \neq i} \delta_x(x_i, x_j) \), where \( \delta_x \) denotes the associated metric of space \( X \).

- For all \( j^* \in J \), find the spread of ranks, i.e. \( r_{i,max} \) and \( r_{i,min} \) of \( y_{j^*} \) in terms of \( \delta_y \) such that:

\[
\begin{align*}
r_{i,max} &= \# \{ j : j \neq i, \delta_y(y_j, y_i) \leq \delta_y(y_{j^*}, y_i) \} \\
r_{i,min} &= \# \{ j : j \neq i, \delta_y(y_j, y_i) < \delta_y(y_{j^*}, y_i) \}
\end{align*}
\]

- For all rank values \( r_{i,min} < r \leq r_{i,max} \), assign:

\[
P(R = r) = P(R = r) + 1/|J| / (r_{i,max} - r_{i,min}) / n
\]

end

- Compute \( C \) as the empirical CDF of \( \{r_1, \ldots, r_n\} \).

- Return \( d \) as the area under \( C \) normalized by \( (n-1) \)

end

where \( P(R = r) \) is defined as \( P(R = r) = \# \{ i : r_i = r \} / n \), and represents the empirical probability of the rank variable. GMA satisfies the properties of a measure of dependence i.e. it is upper and lower bounded, invariant under a general set of transformations (like rotation and scaling) and can be asymmetric. GMA assumes values between 0.5 and 1.

Being parameter-free, it enjoys a unique computational advantage over other approaches. A simple illustration of GMA over the realizations of two random variables can be shown in Figure 3-1.

3.2.1.4 Time series complexity and permutation entropy

Motivation: All of the previously mentioned measures of dependence look at either the realizations of the random variables or their ranks to measure the involved dependence. This approach can be sometimes vulnerable to noise and outliers. Moreover, it is prone to confuse real dependence between the time series with temporal dependence residing in each. We propose to compute dependence based on derived functions that
Suppose $X = [1, 4, 5]$ and $Y = [1, 7, 3]$. Red arrows are used to show the closest point with respect to the $x$-axis and blue arrows are used to show the closest point with respect to the $y$-axis. Here $P(\text{Rank} = 1) = 1/3$ and $P(\text{Rank} = 2) = 2/3$. hence $\text{GMA} = \text{Normalized Area under CDF} = (1/3 + 1)/2 = 0.67$.

measure time series complexity. The addition of an intermediate “layer” prior to dependence computation might not seem useful at first, however working with functions that encompass complexity information about the original time series is appealing since it helps overcome sensitivity to noise and outliers as will be explained in the following sections.

This approach entails a tradeoff between adding a free parameter (as compared to most information theoretical estimators) and robustness to noise/sensitiveness to abrupt changes. One particular aspect that needs scrutiny here is to assess how much information about the data is preserved using these complexity measures. In the case of permutation entropy for instance (will be described in detail shortly), the complexity value is estimated as the entropy derived from the probability distribution of extracted ordinal patterns. Although such approach is able to describe the nonlinear trends in time series, the lost amplitude information when collecting ordinal patterns hinders the reconstruction of the initial data distribution. It is actually hard to think of any preprocessing function that maps the input
data into a symbolic space that permits the easy extraction of complexity quantifiers while allowing an exact retrieval of the initial data properties. We hence proceed according to a method that retains amplitude while still exploiting the advantage of using ordinal patterns to quantify the nonlinear behavior of time series and refer to this approach as Weighted-Permutation Entropy (WPE)

**Time series complexity:** There is little consensus on the definition of a signal’s complexity. Among the different approaches, entropy-based ones are inspired by either nonlinear dynamics [Pincus, 1991] or symbolic dynamics [Bandt and Pompe, 2002; Kurths et al., 1996]. Permutation entropy (PE) has been recently suggested as a complexity measure based on comparing neighboring values of each point and mapping them to ordinal patterns [Bandt and Pompe, 2002]. Using ordinal descriptors is helpful in the sense that it adds immunity to large artifacts occurring with low frequencies. PE is applicable for regular, chaotic, noisy or real-world time series and has been employed in the context of neural [Li et al.], electroencephalographic (EEG) [Bruzzo et al., 2008; Cao et al., 2004; Li et al., 2008; 2007], electrocardiographic (ECG) [Graff and Kaczkowska, 2012; Zhang et al., 2008] and stock market time series [Zunino et al., 2009].

**Permutation entropy:** Consider the time series \( \{x_t\}_{t=1}^T \) and its time-delay embedding representation \( x_j^{m,\tau} = \{x_j, x_{j+\tau}, \ldots, x_{j+(m-1)\tau}\} \) for \( j = 1, 2, \ldots, T - (m - 1)\tau \), where \( m \) and \( \tau \) denote respectively the embedding dimension and time delay. To compute PE, each of the \( N = T - (m - 1)\tau \) subvectors is assigned a single motif out of \( m! \) possible ones (representing all unique orderings of \( m \) different real numbers). PE is then defined as the Shannon entropy of the \( m! \) distinct symbols \( \{\pi_i^{m,\tau}\}_{i=1}^{m!} \), denoted as \( \Pi \):

\[
H(m, \tau) = - \sum_{i: \pi_i^{m,\tau} \in \Pi} p(\pi_i^{m,\tau}) \ln p(\pi_i^{m,\tau})
\]

(3–27)

\( p(\pi_i^{m,\tau}) \) is defined as:

\[
p(\pi_i^{m,\tau}) = \frac{\| \{ j : j \leq N, \ type(x_j^{m,\tau}) = \pi_i^{m,\tau} \} \|}{N}
\]

(3–28)
where \( \text{type}(.). \) denotes the map from pattern space to symbol space and \( \| . \| \) denotes the cardinality of a set. An alternative way of writing \( p(\pi_i^{m,\tau}) \) is:

\[
p(\pi_i^{m,\tau}) = \frac{\sum_{j \leq N} 1_{u:\text{type}(u) = \pi_i(x_j^{m,\tau})}}{\sum_{j \leq N} 1_{u:\text{type}(u) \in \Pi(x_j^{m,\tau})}}
\]

where \( 1_A(u) \) denotes the indicator function of set A defined as \( 1_A(u) = 1 \) if \( u \in A \) and \( 1_A(u) = 0 \) if \( u \notin A \). PE assumes values between in the range \( [0, \ln m!] \) and is invariant under nonlinear monotonic transformations.

### 3.2.2 Generalized Association for Time Series

#### 3.2.2.1 Motivation

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

#### 3.2.2.2 Proposed algorithm

For each realization in the time series, we dismiss the realizations within a neighboring window to discard dependence pertaining to time structure. Only points falling outside that window would be accounted for. The latter might include points revisited by the time series as time passes. The choice of the 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 Algorithm 3 and Figure 3-2 shows an illustrative example:

**Algorithm 3:** Time Series Generalized Measure of Association (TGMA)

**Input:** Bivariate time series \( \{x_t, y_t\}_{t=1}^{n} \) assuming values in the joint space \( X \times Y \)

**Output:** Estimated dependence \( d \in [0.5 : 1] \)

```
begin
Initialization:
- Assign \( P(R = r) = 0 \ \forall \ r \in \{1, \ldots, (n - 1)\} \)
- Let \( \xi_x \) and \( \xi_y \) denote the ACFs of \( x \) and \( y \), and let \( 0 < l^*_x, l^*_y < n \).
- Set \( l^*_x \) as follows:
  \[
  l^*_x = \begin{cases} 
    \text{lag at first zero-crossing of } \xi_x & \text{if } \xi_x \text{ has a zero-crossing.} \\
    \text{lag at first minimum of } \xi_x & \text{if } \xi_x \text{ has no zero-crossing.} \\
    \text{lag at } 1/e \text{ decay of } \xi_x & \text{if } \xi_x \text{ has no zero-crossing & is monot.}
  \end{cases}
  \]
- Set the value of \( l^*_y \) in a similar fashion with respect to \( \xi_y \).

for \( i \in \{1 \ldots n\} \) do
  a. Find \( x_{j^*} \) (\( j^* \in J \)), where \( j^* \) satisfies:
  \[
  j^* = \arg \min_{|j-i| \geq \max(l^*_x, l^*_y)} \delta_x(x_i, x_j)
  \]
  b. For all \( j^* \in J \), find the spread of ranks, i.e. \( r_{i,\text{max}} \) and \( r_{i,\text{min}} \) of \( y_{j^*} \) in terms of \( \delta_y \) such that:
  \[
  r_{i,\text{max}} = \#\{j : j \neq i, \delta_y(y_j, y_i) \leq \delta_y(y_{j^*}, y_i)\}
  \]
  \[
  r_{i,\text{min}} = \#\{j : j \neq i, \delta_y(y_j, y_i) < \delta_y(y_{j^*}, y_i)\}
  \]
  c. For all rank values \( r_{i,\text{min}} < r \leq r_{i,\text{max}} \), assign:
  \[
  P(R = r) = P(R = r) + 1/|J|/(r_{i,\text{max}} - r_{i,\text{min}})/n
  \]
end
- Compute \( C \) as the empirical CDF of \( \{r_1, \ldots, r_n\} \).
- Return \( d \) as the area under \( C \) normalized by \( (n - 1) \)
end
```

A thorough discussion of TGMA within a comparative approach to Kendall’s tau applied on copula generated data and real data can be found in [Fadlallah et al., 2012a].
3.2.3 Weighted-Permutation Entropy Based Dependence

3.2.3.1 Motivation

The main shortcoming in the definition of PE provided in Section 3.2.1.4 resides in the fact that no information besides the order structure is retained when extracting the ordinal patterns for each time series. This may be inconvenient for the following reasons: (i) most time series have information in the amplitude that might be lost when solely extracting the ordinal structure (ii) ordinal patterns where the amplitude differences between the time series points are greater than others should not contribute similarly to the final PE value and (iii) ordinal patterns resulting from small fluctuations in the time series can be due to the effect of noise and should not be weighted uniformly towards the final value of PE. Figure 3-3 shows how the same ordinal pattern can originate from different $m$–dimensional vectors.

We hence suggest a modification that alters the way PE handles the patterns extracted from a given signal. The new scheme better tracks abrupt changes in the signal...
and assigns less complexity to segments that exhibit regularity or are subject to noise effects.

### 3.2.3.2 Weighted-permutation entropy

To counterweight the aforementioned shortcomings, we extend the current PE procedure to incorporate significant information from the time series when retrieving the ordinal patterns. The main motivation lies in saving useful amplitude information carried by the signal. We refer to this procedure as weighted-permutation entropy (WPE) and summarize it in the following steps. First, the weighted relative frequencies for each motif are calculated as follows:

\[
p_w(\pi_i^{m,\tau}) = \frac{\sum_{j \leq N} 1_{u:\text{type}(u) = \pi_i(x_j^{m,\tau}).w_j}}{\sum_{j \leq N} 1_{u:\text{type}(u) \in \Pi(x_j^{m,\tau}).w_j}} \tag{3-30}
\]

WPE is then computed as:

\[
H_w(m, \tau) = - \sum_{i: \pi_i^{m,\tau} \in \Pi} p_w(\pi_i^{m,\tau}) \ln p_w(\pi_i^{m,\tau}) \tag{3-31}
\]
Note that when \( w_j = \beta \ \forall \ j \leq N \) and \( \beta > 0 \), WPE reduces to PE. It is also interesting to highlight the difference between the definition of weighted entropy in this context and previous ones suggested in the literature. Weighted entropy, defined as

\[
H_{we} = -\sum_k w_k p_k \ln p_k,
\]

has been suggested as a variant to entropy that uses a probabilistic experiment whose elementary events are characterized by weights \( w_k \) [Guiasu, 1971]. WPE on the other hand, extends the concept of PE while keeping the same Shannon’s entropy expression reflected by Equation 3–31, hence weights are added prior to computing the \( p(\pi_i^{m,\tau}) \). The choice of weight values \( w_i \) is equivalent to selecting a specific (or combination of) feature(s) from each vector \( x_j^{m,\tau} \). Such features may differ according to the context used. Note that the relation \( \sum_i p_w(\pi_i^{m,\tau}) = 1 \) still holds. In this manuscript, we use the variance or energy of each neighbors vector \( x_j^{m,\tau} \) to compute the weights. Let \( \bar{x}_j^{m,\tau} \) denote the arithmetic mean of \( x_j^{m,\tau} \) or:

\[
\bar{x}_j^{m,\tau} = \frac{1}{m} \sum_{k=1}^{m} x_{j+(k+1)\tau}
\]

(3–32)

We can hence express each weight values as:

\[
w_j = \frac{1}{m} \sum_{k=1}^{m} \left( x_{j+(k-1)\tau} - \bar{x}_j^{m,\tau} \right)^2
\]

(3–33)

The motivation behind this setting is to specifically counteract the limitations discussed in the previous section, i.e. weight differently neighboring vectors having the same ordinal patterns but different amplitude variations. In this way, WPE can be also used to detect abrupt changes in noisy or multicomponent signals. The modified \( p(\pi_i) \) can be then thought of as the proportion of variance accounted for by each motif. The above definition of WPE retains most of PE’s properties and is invariant under affine linear transformations. WPE however presents a specificity, given it incorporates amplitude information and demonstrates more robustness to noise. A more thorough discussion of WPE can be found in [Fadlallah et al., 2013a].
3.2.3.3 WPE-based dependence

Generally speaking, every time series is assigned a scalar value for PE and WPE.
We propose to slide a window on each time series to generate “complexity” curves whose
length is determined according to the chosen window length. Once we have these curves,
we can apply any measure of association such that Spearman’s rho, Kendall’s tau or GMA
to capture the monotonicity of the dependence between the two curves. More elaboration
on the use of WPE as a novel complexity measure for time series, that incorporates
amplitude information can be found in Appendix D.
CHAPTER 4
THE DYNAMIC DEPENDENCE GRAPH

Graph theoretical methods have been already proposed as a tool to analyze structural, functional and effective brain connectivity [Bullmore and Sporns, 2009], where vertices correspond to brain regions or neurons, and edges represent synapses or paths of pronounced statistical association between neural elements. Dependence graphs have proved their usefulness in describing dependence relations between random variables and time series [Lauritzen, 1996]. In this chapter, we show how to use graph theoretical concepts to tackle the problem formulated in Section 1.1.6.

4.1 Graph Model

We model the electrodes network as a complete undirected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges [Fadlallah et al., 2013b]. For each edge $e_{ij}$ between two vertices $i$ and $j$, we assign a value $m_{ij}$ representing the dependence between the processed signals recorded at the electrodes locations. Note that for correlation, $m_{ij}$ exists in the interval $[-1, 1]$, whereas for GMA it lies in the interval $[0.5, 1]$. We assume that the graph is undirected since our goal is to quantify dependencies between brain regions, which does not account for direction. To relax this assumption, measures of causality and a subsequent directed graph can be considered for assessing effective connectivity. The graph $G$ is further transformed into an incomplete graph by discarding values falling below a predefined statistical threshold (more details in Section 4.3). As a result, and after averaging over trials, an $N_c \times N_c \times N_w$ adjacency matrix $M_d^c$ is generated for each dependence measure $d$ and condition $c$, where $N_w$ refers to the number of frames or time windows.

Choosing the dependence measure is crucial for our approach. Correlation can be used in this regard, but since it only captures second order interactions, it performs rather poorly for time series featuring higher order interactions. A thorough discussion of possible
measures of dependence (or causality) that can be used to weight the graph edges can be found in Chapter 3.

4.2 Graph Theoretical Concepts

4.2.1 Basic Notations

The neighborhood of a vertex $v \in V$ is the set $\mathcal{N}_v$ of all vertices (or neighbors) connected to $v$, i.e $\mathcal{N}_v = \{r : e_{rv} \in E \text{ or } m_{rv} \neq 0\}$. The degree of $v$ consists of the number of vertices that are incident to $v$, i.e. $n_v = |\mathcal{N}_v|$, where $|.|$ denotes the cardinality of a set. A path from vertex $r$ to vertex $s$ is a sequence of vertices and edges that begins with $r$ and ends with $s$, with an edge connecting each vertex with the succeeding one. The distance $d_{r,s}$ between $r$ and $s$ is the minimum length of any path connecting the two vertices. We hereby present some graph theoretical measures of interest for our analysis.

4.2.2 Node Clustering Coefficient

The clustering coefficient is a frequently used measure to characterize the local and global structure of unweighted graphs [Boccaletti et al., 2006]. The clustering coefficient measures the extent to which nodes in a graph tend to cluster together. It is defined for a node $i$ in the graph as:

$$CC_i = \frac{|\{e_{jk} : v_j, v_k \in \mathcal{N}_i, e_{jk} \in E\}|}{n_i(n_i - 1)} \quad (4-1)$$

where $n_i = |\mathcal{N}_i|$. The clustering coefficient assumes values between 0 and 1. A common practice is to average the clustering coefficients of all nodes in the graph:

$$CC = \frac{1}{|V|} \sum_{i=1}^{n} C_i \quad (4-2)$$

The clustering coefficient of a vertex can be interpreted in terms of its tendency to promote connections among its neighborhood, and can therefore be considered as an indicator of information flow in dynamic networks [López-Fernández et al., 2006].
4.2.3 Measures of Centrality

A vertex’s importance in a graph can be quantified using several measures, including measures of centrality.

**Betweenness Centrality:** The betweenness centrality of a vertex \( i \) measures the number of shortest paths traversing that vertex. It was first proposed by Anthonisse [1971] and has been later used in several contexts [Barthelemy, 2004; Freeman, 1977; Goh et al., 2003]. It can be defined as:

\[
B_i = \sum_{j \neq i \neq k} \frac{\text{No. shortest paths from } j \text{ to } k \text{ via } i}{\text{No. shortest paths from } j \text{ to } k} \tag{4-3}
\]

Alternatively:

\[
B_i = \sum_{j \neq i \neq k} \frac{\sigma_{jk}^i}{\sigma_{jk}} \tag{4-4}
\]

where, in Equation 4–4, \( \sigma_{jk}^i \) denotes the number of shortest paths from \( j \) to \( k \) including \( i \). When the betweenness centrality of a vertex is high, the vertex is more likely to be an intermediate communication node in the graph. Such vertices can be seen as occupying the “structural holes” in the network [Burt, 1995; López-Fernández et al., 2006].

**Subgraph Centrality:** The subgraph centrality of a vertex \( i \) can be defined as the weighted sum of closed walks having different lengths, starting and ending at \( i \).

\[
S_i = \sum_{k=0}^{\infty} \frac{\mu_i^k}{k!} \tag{4-5}
\]

where in Equation 4–5, \( \mu_i^k \) refers to the \( k^{th} \) local spectral moment, which defines the number of closed walks of length \( k \), starting and ending on \( i \). \( \mu_i^k \) is computed using the \( i^{th} \) diagonal entry of the \( k^{th} \) power of the graph adjacency matrix.

\[
\mu_i^k = (M_d^k)_{ii} \tag{4-6}
\]
Due to space limitations, the computational details pertaining to subgraph centrality are omitted and have been thoroughly discussed by Estrada and Rodriguez-Velazquez [2005].

**Closeness Centrality:** The closeness centrality of a vertex $i$ is defined as the inverse of the sum of distances between $i$ and all other vertices.

\[
C_i = \frac{1}{\sum_{j \in V} d_{i,j}} \quad (4-7)
\]

### 4.2.4 Local Efficiency

The local efficiency of a vertex $i$ can be defined in terms of the sum of inverse distances between the vertices in $N_i$. The local efficiency mainly measures the efficiency in communication between the direct neighbors of $i$, when the node itself is removed.

\[
E_{i}^{loc} = \frac{1}{n_i(n_i - 1)} \sum_{j,k \in N_i} \frac{1}{d_{j,k}} \quad (4-8)
\]

### 4.2.5 Connected Components

A connected component of an undirected graph is a maximal subgraph in which any two vertices are connected to each other by paths. The concept is graphically illustrated in Figure 4-1.

![Figure 4-1](image)

Figure 4-1. An undirected graph consisting of three connected components.
4.2.6 Graph Visualization

Visualizing the pairwise values of dependence for 129 or 257 sensors can be a challenging task. To increase the user friendliness of the graph displays, we map the EEG sensors to the circumference of a circle, in a way that preserves the main structures of the scalp geography. This is illustrated in Figure 4-2.

![Map to graph vertices](image)

Figure 4-2. The standard 10-20 configuration (Refer to Figures 2-5 and 2-6) is used to map the EEG sensors to the circumference of a circle. The circle still retains the main structures of the scalp geography in the sense that lower vertices correspond to the occipital areas, whereas the higher ones correspond to frontal areas.

4.3 Weights Computations

For each measure of dependence, we define $M_d$ as the difference between matrices pertaining to each condition, i.e. $M_d = M_d^F - M_d^G$. It is then possible to visualize how pairwise dependences vary in sensor space across time. The resulting matrices are first
processed for statistical significance by discarding all values falling within 2 standard deviations of their means.

4.3.1 Time-Delay Embedding

Prior to computing some measures of dependence (like mutual information or GMA), we perform a time-delay embedding for each time series $x^{(i,k)}$ and $y^{(i,k)}$. The motivation behind this is to work on a reconstructed state-space version of the dynamical system that would account for propagation delays, especially between neighboring channels. We further suggest a method based on synthetic data to analyze the embedding parameters when using these measures (Section 4.3.2). Again, for each time series, we define an embedding $\xi^{(i,k)}_{j}(m, \tau)$ as:

$$\xi^{(i,k)}_{j}(m, \tau) = (x^{(i,k)}_{j}, x^{(i,k)}_{j+\tau}, \ldots, x^{(i,k)}_{j+(m-1)\tau}) \quad (4-9)$$

for $j = 1, 2, \ldots, (N_s - N_r) - (m - 1)\tau$, where in the above equation, $\tau$ and $m$ denote the time delay and embedding dimension and $\xi^{(i,k)}_{j}(m, \gamma)$ the $j^{th}$ reconstructed vector with time delay $\tau$ and embedding dimension $m$. Similar embedding is performed for each $y^{(i,k)}$.

In our computations, we used time windows of 114 ms corresponding to 114 samples given $d = 1$ and $F_s = 1000$ Hz. The first duration corresponds to two cycles of a sinusoid with frequency $F_o$ and covers roughly the propagation time from the occipital to the frontal cortices. Time series were embedded in $\gamma = 8$ dimensions to account for the propagation delay among neighboring channels when computing certain measures of dependence such as mutual information.

4.3.2 Impact of Free Parameters

4.3.2.1 Embedding dimension

Embedding in time is a crucial step because of the limitation of the number of samples. The chosen value for the embedding dimension $\gamma$ should not be too high for computational constraints and not too low because this undermines the goal of capturing the delayed dependence. To estimate $\gamma$, simulations were performed on two synthetically
generated noisy time series. Each consisted of a two-period sinusoid oscillating at a frequency $F_0$, corresponding to 114 samples, with additive white gaussian noise (AWGN) in the phase and amplitude. One of the time series was shifted by 25 samples with respect to the other and our goal was to estimate which embedding dimension (or time lag for correlation) achieves a level of dependence equal to the one recorded between the two non-shifted time series. Specifically, $s_1(t) = S_1 \sin(2\pi F_0 t + \mathcal{N}(\mu_1, \sigma_1^2)) + \mathcal{N}(\mu_2, \sigma_2^2)$ and $s_2(t) = S_2 \sin(2\pi F_0 (t + \delta_{s_1-s_2}) + \mathcal{N}(\mu_1, \sigma_1^2)) + \mathcal{N}(\mu_2, \sigma_2^2)$ where $\delta_{s_1-s_2} = 25$, $t = 1/F_0[1, \ldots, (114 + \delta_{s_1-s_2})]$. $S_1, S_2$ were chosen to be 5 and 3 and $\mu_1, \mu_2, \sigma_1, \sigma_2$ were respectively set to 0.1, 0.15, 0.3 and 0.4. Figure 4-3 (c and d) shows that for GMA and MI, an embedding dimension close to 8 achieves the desired level of dependence. Besides, as anticipated, a time lag of 25 samples was needed to maximize the correlation between the two time series. Hence, to maximize correlation when computing dependence on the filtered time series, we iterate over lags corresponding to one period of the signal or 54 samples and when computing absolute correlation, we simplify the problem by looking at lags between 1 and 29 samples (the equivalent of a half-period accounting for out of phase oscillations).

#### 4.3.2.2 Time delay

We use the same simulation setting described above to analyze the impact of the time delay $\tau$. Values for $\tau$ ranging between 1 and 5 were utilized to embed the two time series. Figure 4-3 (e) shows that the impact of $\tau$ is not as important as that of the embedding dimension $\gamma$ since the dependence curves do not change much when increasing the time delay value.

#### 4.3.2.3 Time window

Time windows of 114 and 456 samples were used for computing dependence values. While the first one achieves a higher time resolution, the second has more samples and yields a more reliable estimator of the joint density.
4.3.2.4 Time complexity

The time complexity for the mutual information estimator is $O(n^2\gamma)$ assuming a straightforward neighbour search implementation. GMA also runs in $O(n^2\gamma)$ since a rank computation step is performed for each point of the time series. In our simulations, GMA was noticeably faster than mutual information but obviously both performed slower than Pearson’s correlation.

![Figure 4-3.](image)

(a) Synthetic data: two noisy time series $s_1(t)$ and $s_2(t)$ oscillating at the same frequency with different amplitudes. $s_2(t)$ has been delayed by 25 samples. (b) Scatter plot of the two time series. (c) MI value for different embedding dimensions. Red line corresponds to the MI value computed between non-shifted versions of the time series. (d) GMA and correlation values for different embedding dimensions or lags. Blue (red) line corresponds to expected GMA (correlation) value. (e) Plotting GMA values for different time delays does not alter the shape of the curve.
CHAPTER 5
INFORMATION THEORETIC CLASSIFICATION OF DEPENDENCE MATRICES

In this chapter, we discuss how information theoretical concepts can be used to classify cognitive conditions from dependence matrices. We first describe the steps involved in estimating entropy-like quantities using positive definite and infinitely divisible matrices (Refer to Appendix E). We then proceed to creating a null hypothesis of distribution equality based on the estimated entropy values. This null hypothesis can be used in classification context to determine whether a given trial belongs or not to a specific condition, using a maximum likelihood (ML) or maximum a posteriori probability (MAP) approach.

5.1 Preliminary Concepts

Using information theoretic quantities as test statistics is a challenging problem. Operational quantities in Information Theory are typically defined in terms of probability laws underlying the data, which represents a serious limitation since the latter are rarely, if ever, known in most real-world statistical learning settings. In such scenarios, the only source of information is what can be derived from the observed samples; there is a need for novel methods that can be employed as a way of bypassing the density estimation problem.

A current effort in our lab focuses on deriving entropy-like quantities directly from empirical data [Sanchez-Giraldo and Principe, 2013]. This derivation method uses positive definite matrices as suitable descriptors of data and does not assume that probabilities of events are known or have been estimated. As such, the designed functional uses the representational power of positive definite matrices and is typically computed using Gram matrices constructed from pairwise computations between data samples. A quantity similar to mutual information can then be derived using the axiomatic characterization of Renyi’s entropy and the Hadamard product of matrices.

In this chapter, we propose using these information theoretic estimators in the context of cognitive task analysis. Dependence matrices are first constructed from collected and
processed electroencephalographic (EEG) data. We then use matrix entropy estimators to obtain scalar descriptors of these matrices and based on these quantities, estimate mutual information between any two matrices, in order to determine whether they belong to the same or distinct cognitive conditions.

The rest of the chapter is organized as follows. In Section 5.2, we briefly discuss the mathematical foundations of the information theoretic framework. Section 5.3 describes the preprocessing done on the dependence matrices and in Section 5.4, we describe the computational steps involved in computing network specific quantities from the derived dependence matrices. Sections 5.5 and 5.6 provide an overview of the results and formulate the problem as a classification problem, and Section 5.7 offers discussion and concluding remarks.

5.2 Estimating Entropy-Like Quantities with Positive Definite Matrices

5.2.1 Matrix Entropy Estimator

Let \( M_n \) denote the algebra of \( n \times n \) matrices over real numbers. Consider the set \( \Delta^+_n \) of positive definite matrices in \( M_n \) for which \( \text{tr}(A) \leq 1 \), and let \( A \in \Delta^+_n \) and \( B \in \Delta^+_n \). When \( \text{tr}(A) = \text{tr}(B) = 1 \), the functional:

\[
S_\alpha(A) = \frac{1}{1-\alpha} \log_2 (\text{tr}(A^\alpha)),
\]

(5–1)
satisfies the following set of conditions (refer to [Sanchez-Giraldo et al., 2012] for proofs):

1. \( S_\alpha(PAP^*) = S_\alpha(A) \) for any orthonormal \( P \in M_n \).
2. \( S_\alpha(pA) \) is a continuous function for \( 0 < p \leq 1 \).
3. \( S_\alpha(\frac{1}{n}I) = \log_2 n \) (entropy is exhaustive).
4. \( S_\alpha(A \otimes B) = S_\alpha(A) + S_\alpha(B) \).
5. If \( AB = BA = 0 \), then for the strictly monotonic and continuous function \( g(x) = 2^{(\alpha-1)x} \) for \( \alpha \neq 1 \) and \( \alpha > 0 \), we have that:

\[
S_\alpha(tA + (1-t)B) = g^{-1} (tg(S_\alpha(A)) + (1-t)g(S_\alpha(B)))\). 
\]

(5–2)
The computed $S_\alpha(A)$ can be seen as an analogous quantity to Renyi’s entropy, although no estimation of PDFs was undertaken. This is possible because there is an axiomatic formulation of entropy, therefore the goal of this development is to create operators in reproducing kernel Hilbert spaces (RKHSs) that display the same functional properties. The next natural step would be then to define a quantity similar to mutual information, and to achieve that, a term that represents joint entropy needs to be derived.

### 5.2.2 Matrices Joint Entropy Estimator using Hadamard Product

The Hadamard product between matrices $(\odot)$ can be used to define a quantity that is analogous to joint entropy as follows:

$$
S_\alpha\left(\frac{A \odot B}{\text{tr}(A \odot B)}\right) = \frac{1}{1 - \alpha} \log_2 \left( \text{tr}\left(\frac{A \odot B}{\text{tr}(A \odot B)}\right)^{\alpha}\right)
$$

(5–3)

Using Equation 5–3, the expression defining mutual information can be then written as:

$$
I_\alpha(A, B) = S_\alpha(A) + S_\alpha(B) - S_\alpha\left(\frac{A \odot B}{\text{tr}(A \odot B)}\right),
$$

(5–4)

where in Equation 5–4, $A$ and $B$ are positive semidefinite matrices with nonnegative entries and unit trace such that $A_{ii} = \frac{1}{n}$ for all $i = 1, \ldots, n$. Notice that, when defined this way, the above quantity is nonnegative and satisfies:

$$
S_\alpha(A) \geq I_\alpha(A, A).
$$

### 5.3 The Dependence Matrices

In the remainder of this chapter, and for mathematical notation brevity, we refer to the face condition by the superscript $F$ and the Gabor patch condition by the superscript $G$. Three rank-based nonparametric measures of statistical dependence were used in the calculations. The first two are traditional measures of correlation included for the sake of comparison: Spearman’s rho and Kendall’s tau. The third is a novel measure of association that can capture nonlinear interactions while attenuating the effect of temporal dependence: Time Series Generalized Measure of Association (TGMA). Typically, when
considering realizations from two time series, the nearest neighbor in amplitude for a
given point is simply the nearest in time. Importantly, this temporal contiguity does
not reveal the dependence structure in the data. TGMA tries to address this problem
by reducing the contribution of neighbors that are very close in time. The algorithm’s
computational steps have been already outlined in Chapter 3 and Algorithm 3 and a
thorough discussion can be found in Fadlallah et al. [2012a]. An important advantage
of the TGMA approach to association that is not exploited in this dissertation is that it
allows computing dependence between variables with different metric spaces.

We symmetrize the values of TGMA by averaging forward and backward dependence
measures. Therefore for each pair of channels $i$ and $j$, we assign a value $m_{ij}$ representing
the dependence between the processed signals recorded at the electrodes locations. For
Spearman’s rho and Kendall’s tau, $m_{ij}$ exists in the interval $[0, 1]$, whereas for TGMA it
lies in the interval $[0.5, 1]$. As a result, and after averaging over trials, an $N_c \times N_c \times N_w$
dependence matrix $M_{d,k}^c$ is generated for each dependence measure $d$, trial $k$ and condition
c, where $N_c = 129$ or $N_c = 257$ depending on the data acquisition system and $N_w = 39$
refers to the number of frames or time windows over which computations were carried out.

5.4 Computational Steps

Let $M_{d,k}^F$ and $M_{d,k}^G$ be the dependence matrices corresponding to the two conditions F
(neutral face) and G (Gabor patch). For simplicity, without loss of generality, we average
the matrices over the time windows dimension. We perform two sets of computations. In
the first, we average the matrices over all trials and then compute their corresponding
$I_{\alpha}(\ldots)$ value as defined in Equation 5–4. In the second, we perform the same computa-
tions, this time using all possible combinations of trials and using both conditions to serve
as a surrogate statistic. Computations are performed for matrices belonging to the same
and different conditions. The logic underlying this paradigm is to verify whether we do
obtain different distributions when using matrices belonging to different visual perception
conditions. The key steps are outlined below:
1. Linearly scale the TGMA matrices to the same range of Kendall’s tau and Spearman’s rho. While this step is not mandatory, it is still desirable to preclude any systematic differences due to the range of the input data. At the same time, it also enables performance evaluation across different measures of dependence. Scaling the matrices can be done using:

\[
M_{\text{TGMA},k}^c \leftarrow 2M_{\text{TGMA},k}^c - 1.
\]

2. Ensure positive definiteness of the matrices (by checking the eigenvalues or attempting a Cholesky decomposition) and their infinite divisibility. To motivate this step, it is useful to remind that the functional \( S_\alpha \) defined in Equation 5–1 is applied on Gram matrices constructed from pairwise evaluations of a positive definite kernel. However, the matrices of dependence do not necessarily represent Gram matrices and might not be infinitely divisible. This is because the condition

\[
\sum_{i=0}^{n} \sum_{j=0}^{n} \alpha_i \alpha_j d^2(x_i, x_j) \leq 0,
\]

for any \( \alpha \in \mathbb{R}^{n+1} \) with \( \sum_{i=0}^{n} \alpha_i = 0 \) might not always be met when \( d(x_i, x_j) \) is derived from the association between the time series \( x_i \) and \( x_j \).

Note that TGMA (and measures of absolute correlation) can be easily framed in the form of as a distance measure using:

\[
M_{\text{TGMA},k}^c \leftarrow 1 - M_{\text{TGMA},k}^c
\]

(5–5)

The definition in Equation 5–5 does not necessarily satisfy the fourth characteristic of a distance function \(^1\), which is the triangle inequality, and as a result does not

---

\(^1\) A function \( d \) defined on \( \mathbb{X} \) is a distance if for \( x, y, z \in \mathbb{X} \), the below conditions are met:

1. \( d(x, y) \geq 0 \)
2. \( d(x, y) = 0 \) if \( x = y \)
3. \( d(x, y) = d(y, x) \)
4. \( d(x, z) \leq d(x, y) + d(y, z) \)
necessarily entail a distance. Therefore, we check that \( \exp(-rd^2(x_i, x_j)) \) defines a positive definite matrix because matrices derived from such functions would be infinitely divisible (Refer to Appendix E). This can be seen as well from a different perspective. Proposition E.2.1 states that \( B = -\log A \) for an infinitely divisible matrix \( A \) and a negative definite matrix \( B \). So \( A = \exp C \) for the positive definite matrix \( C = -B \), and this is inline with the computation of \( \exp(-rd^2(x_i, x_j)) \).

3. Normalize the matrices traces to 1, as per the suggestion of Sanchez-Giraldo and Principe [2013]. This can be done using:

\[
M_{d,k}^c(i, j) \leftarrow \frac{M_{d,k}^c(i, j)}{\sqrt{M_{d,k}^c(i, i)M_{d,k}^c(j, j)}}.
\]

4. Set the value of \( \alpha \) to 1.01 as per the suggestion of Sanchez-Giraldo and Principe [2013]. Subsequently, for each measure of dependence, perform the following sets of computations:

[Using averaged trials]

(a) Average matrices over trials to obtain \( M_d^F \) & \( M_d^G \).

(b) Compute the following quantity:

\[
Q_{avg}^{FG} = I_\alpha(M_d^F, M_d^G).
\]

[Using all the trials]

(a) For all different pairs of trials \( m \) and \( n \) belonging to the same condition (say \( F \)), compute:

\[
Q_{m,n}^F = I_\alpha(M_{d,m}^F, M_{d,n}^F).
\]

(b) Repeat the same for pairs of trials belonging to two different conditions, and compute
5. Check which dependence measure achieves the least value of $Q_{avg}^{FG}$.

6. Compare the surrogate distributions separability of $Q_{avg}^{FG}$ and $Q^F$ for each measure of dependence.

5.5 Results

As mentioned in Section 2.1, the dimensionality of the dependence matrices varied from $129 \times 129$ to $257 \times 257$ between the first and last participants, which means that each dependence matrix consisted of 12964 or 32896 entries. A total of 40 trials were performed per participant, equally divided between the two conditions (20 trials for F and 20 trials for G). Figure 5-1 shows the obtained mean values of $Q_{avg}^{FG}$ (as expressed in Equation 5–6) for each of the three measures, with the corresponding variance across participants.

We then generate the distribution of $I_\alpha(A, B)$, when $A$ and $B$ correspond to trials from the same condition, or different conditions. In both cases, all possible combinations of trials are considered. When $A$ and $B$ correspond to the same condition, the total number of possible combinations is $20 \times 19 = 380$, because there is no need to compute $I_\alpha(A, A)$, and when $A$ and $B$ belong to different conditions, the total number of possible combinations is $20 \times 20 = 400$. Figure 5-2 shows the obtained histograms of $Q_{m,n}^{FG}$ and $Q_{m,n}^F$ (as expressed in Equations 5–7 and 5–8) for each of the three measures. The obtained distributions (Figure 5-2) are congruent with our expectations, since we anticipated that two matrices belonging to different conditions would achieve a lesser level of mutual information. This is exactly what we observe. The reason why we initially expected a lower mutual information for different conditions is that mutual information can be interpreted as the information gain from assuming independence to knowing the joint distribution. For different conditions, no substantial information gain is achieved between assuming independence and knowing the joint distribution because it is easier to express
Figure 5-1. Computed values of $Q_{\text{avg}}^{FG}$ for TGMA, Kendall’s tau and Spearman’s rho. For each participant and condition, the dependence matrices were averaged over trials and time windows. The errorbars correspond to one standard deviation of the mean across participants.

the latter as a product of the marginals, as compared to a case in which the matrices belong to the same condition, where such an assumption would be totally misleading.

5.6 Classification

A similar approach to the one described in the previous section can be adopted for supervised classification. We can train with data gathered from 5 participants to generate distributions of entropy estimates for the two conditions, then use the $6^{th}$ participant for testing. With 40 trials per subject, the number of training matrix samples amounts to 100 per each condition. For the first 5 subjects, we label each dependence matrix with its condition (after averaging over time windows) and therefore obtain one distribution of entropy values per condition. The 20 trials belonging to the $6^{th}$ subject are then assigned
Figure 5-2. Distributions of $Q_{m,n}^{FG}$ and $Q_{m,n}^F$ for TGMA, Kendall’s tau and Spearman’s rho. Each distribution corresponds to mutual information estimates from trials belonging to the same (or separate) conditions. 380 estimates of mutual information (all unique combinations of 20 face trials) were obtained for trials belonging to the same condition (green distributions). 400 estimates of mutual information (all distinct combinations of 20 face trials and 20 Gabor patch trials) were obtained for trials belonging to different conditions (red distributions).

labels based on where they fall in the distributions. The classification results using this approach are shown in the Results Chapter (Chapter 6, Table 6-5).
5.7 Conclusions

In this chapter, we demonstrated that it is possible to use information theoretic concepts to characterize distinct visual cognitive states by using dependence matrices extracted from processed human EEG data. Dependence between pairwise channels was computed with two traditional (Kendall’s tau and Spearman’s rho) and one novel (TGMA) rank-based estimator. Entropy-like quantities were then estimated from the dependence matrices. Positive definiteness enables defining analogues to Renyi’s marginal and joint entropies, and therefore provides a means to estimate mutual information between matrices. Two sets of computations were performed: in the first, we averaged over trials and calculated the corresponding mutual information between matrices belonging to two different conditions and in the second, we used all trials to generate distributions of condition equality and inequality. Computations were carried out using a carefully selected value of the free parameter $\alpha$ ($\alpha = 1.01$), as per the suggestion of Sanchez-Giraldo and Principe [2013]. Trying other values did not seem to significantly alter the results.

Results revealed that a lower value of mutual information is achieved across participants for TGMA. All methods generated separable distributions between the two visual perceptual conditions with overall higher discriminative power observed for the TGMA algorithm.
CHAPTER 6
COMPUTATIONS AND RESULTS

6.1 Overview and Computational Settings

This chapter delves into more details about the computational procedures and presents the main results. The results presentation will be organized as follows:

1. **Results with respect to a reference channel**: In this section, dependence values computed with respect to a reference channel are projected onto sensor space and displayed over the scalp.

2. **Results using all the channels**: In this section, all pairwise values of dependence are evaluated, then visualized in a graph that evolves with time. To further understand the network structure in the graphs, we extract two types of features describing the graph characteristics:
   (a) **Local descriptors** of the nodes using graph theoretical quantities. The extracted features include node degree, clustering coefficient, betweenness centrality, subgraph centrality, and local efficiency, and are again projected onto the scalp.
   (b) **Global descriptors** of the graph structures. Estimates of matrix entropies or mutual information between dependence matrices belonging the the same or different conditions, can be used to generate distributions that are later used for classification.

3. **Consistency results**: In this section, we compute Cronbach’s alpha per condition for all the participants. To get a better idea about consistency across subjects, Cronbach’s alpha is also computed after concatenating trials from all subjects.

4. **Performance results**: In this section, we assess the methods performance using two approaches:
   (a) Applying the KS test for all subjects. To do so, we average the matrices over time windows and trials, per condition.
Learning from the data and evaluating the classification performance. We train with 5 subjects (100 trials per condition) and classify the trials pertaining to the 6th subject (20 trials per condition).

Dependencies are computed per trial for every pair of channels. For each dependence measure and condition, an $N_c \times N_c \times N_t \times N_w$ matrix was constructed, where $N_w$ denotes the number of time windows over which dependence is computed. For $d = 1$ and a time window length of 114 ms, $N_w = 36$. Except for cases where we are interested in the time evolution of dependence, we choose to average the matrices of dependence over the time windows. Most results were visualized using scalp projections [EMEGS, 2011], or dependence graphs. The following measures of dependence were used to weight the edges of the dependence graph:

1. Pearson’s correlation coefficient [Pearson, 1896]
2. Mutual information using a kNN approach [Kraskov et al., 2004]
3. Spearman’s rank correlation coefficient [Spearman, 1904]
4. Kendall’s rank correlation coefficient [Kendall, 1938]
5. Generalized Measure of Association (GMA) [Seth, 2011]
6. Time Series Generalized Measure of Association (TGMA) [Fadlallah et al., 2012a]
7. Permutation Entropy (PE) [Bandt and Pompe, 2002]
8. Weighted Permutation Entropy (WPE) [Fadlallah et al., 2013a]
9. Phase Synchrony or Phase Locking Value (PLV) [Wan et al., 2013]

Absolute correlation was taken into consideration to achieve a better comparison with the other dependence measures since negative correlation values with high magnitude indicate anticorrelation and hence correspond to regions of strong statistical dependence. Furthermore, for measures 2, 4 and 5, signals were embedded in $m = 8$ dimensions as per Section 4.3.2. To account for a similar effect in correlation, dependence values were maximized over time lags between 1 and 29 samples. As a result, a dependence map could
be drawn for each channel, showing sensor locations exhibiting high levels of dependence, in heat map format.

6.2 The Kolmogorov-Smirnov Test

To assess quantitatively the dependency maps between conditions, we use the two-sample Kolmogorov-Smirnov (KS) test, which is a non-parametric test that compares two sample vectors. The KS test tries to estimate the distance between the empirical distribution functions of the two sets of samples. The null hypothesis is that both samples are drawn from the same distribution. Assuming $\beta_1(x)$ and $\beta_2(x)$ to be the sample vectors, the KS statistic can be calculated as:

$$ KS_{\beta_1, \beta_2} = \max_x |F_{\beta_2}(x) - F_{\beta_1}(x)| $$

(6–1)

where $F_{\beta_1}(x)$ and $F_{\beta_2}(x)$ denote the empirical cumulative distribution functions for the $n$ iid observations, alternatively $F_{\{X_1, \ldots, X_n\}}(x) = \frac{1}{n} \sum_{i=1}^{n} I_{X_i \leq x}$, where $I_k$ denotes the indicator function.

The null hypothesis is rejected at the $\alpha$-level if $\sqrt{(n_1n_2)/(n_1 + n_2)}KS_{\beta_1, \beta_2} > K_{\alpha}$, where $n_1$ and $n_2$ denote the number of samples from each observation vector and $K$ refers to the Kolmogorov distribution [Marsaglia et al., 2003]. In our case, $n_1 = n_2 = N_t$.

6.3 Cronbach’s alpha

Cronbach’s alpha is a measure of internal consistency, commonly used as an estimate of the reliability of a psychometric test. Cronbach’s alpha was proposed by Cronbach [1951] and may be seen as an indicative of reproducibility of results for a given experimental paradigm.

$$ \alpha = \frac{N_t}{N_t - 1} \left( 1 - \frac{\sum_{i=1}^{N_t} \sigma_i^2}{\sigma_T^2} \right) $$

(6–2)

where, in Equation 6–2, $N_t$ refers to the number of considered items (here $N_t = 15$ or $N_t = 20$, i.e. the number of trials). $\sigma_i$ and $\sigma_T$ denote respectively the variance of the
i\text{th} \text{ item} \text{ and \ the \ variance \ of \ the \ total \ score \ formed \ by \ summing \ all \ the \ items} \ [\text{Bland \ and} \ \text{Altman,} \ 1997].

6.4 \text{ Results}

6.4.1 \text{ Single Reference Channel}

For visualization purposes, we select a channel location in the occipito-parietal region (POz), as shown in Figure 6-1, to display a representative map of the obtained dependencies [Fadlallah \text{ et al.,} \ 2012b]. Two criteria were considered when selecting the channel location, (i) its unbiasedness towards either of the right or left brain hemispheres and (ii) its proximity to the occipital regions where the signal is known to originate from. The advantage sought by selecting this channel consists in providing an insight about how the occipital region of the brain relates in time to the surrounding occipito-temporal and parietal areas.

Figure 6-2 plots dependence maps over the sensor network for correlation, mutual information and GMA. Figures 6-3 and 6-4 show the averaged phase-locking values and their distribution per condition, and each of Figures 6-5, 6-6, 6-7 and 6-8 shows projected dependencies on the right and left lateral parts of the head for the two conditions. GMA was used for Figures 6-6 and 6-7; absolute correlation and MI were used respectively for Figures 6-5 and 6-8. Given its asymmetry, GMA plots were comprised of two rows to reflect both directions of the measure. The lower row corresponds to dependence values computed from all channel locations to channel POz or 72.

\textbf{Correlation:} Absolute correlation shows consistency in locations of active regions across time and conditions. In addition to the visual cortex, these regions seem to point towards sources in the superior temporal sulcus region of the brain.

\textbf{GMA:} Unlike correlation, GMA shows higher dependence for the face condition in the right parietal-temporal-occipital region neighboring P4 (Figures 6-6 and 6-7). This might be explained by a communication between sources in the primary visual cortex and
Figure 6-1. Sensor network and channel 72 highlighted in red. The location of this central electrode is close to the occipital regions where the signal is known to originate from. Ventral region sources and is consistent with engagement of the fusiform or right occipito-temporal areas. A remarkable observation is that this active region seems to be reinforced with the duration of the presentation. The averaged value of GMA in that region shows consistently slight increases with time, noticeable as long as the stimulus is applied, which suggests the enforcement of communication between sources as time passes.

**Mutual Information**: MI tends to show a more balanced distribution of active regions between the two brain hemispheres as compared to correlation and GMA. Some patterns of right parietal activity can be spotted for the face condition, but their corresponding dependence levels vary across time windows. Similar to the two other dependence measures, a slightly higher level of dependence was observed for the face condition.
Figure 6-2. Dependence maps displayed in sensor space in heat map format. Red areas denote regions of higher synchronization with the reference channel (POz or 72), and blue areas denote those that exhibit lower synchronization with the reference channel. Three measures of dependence are illustrated. Top: GMA, Middle: MI and Bottom: Absolute correlation.
**Phase Synchrony**: Figures 6-3 and 6-4 show the projection of phase-locking values over the scalp and the PLVs distribution per condition. Phase synchrony was evaluated on Subject 1 and all the time series samples were used in the computations. Phase synchrony seems to outperform the other dependence measures in discriminability. This however comes at the expense of the reduced time resolution. Theoretically, we expect signals recorded for the face stimulus to be more synchronized than those of the Gabor patch condition, since more structures would be engaged when responding to a facial visual stimulus. Since the cumulative distribution function of the PLVs for the Gabor patch condition increases faster than that for face condition at the smaller PLV range, and slower at the larger PLV range, the PLVs between channels for the Gabor patch stimulus condition is mainly distributed at a small numeric range. Thus, the signals recorded for the Gabor patch condition are less synchronized than those for the face stimulus, which confirms our expectations.

**Dependence across Time Windows**: A subset of electrodes corresponding to the region neighboring electrode location $P_4$ (as shown in Figure 6-9.a) was selected to study the variations of our dependence measures across time windows. GMA values exhibit low variability with slight increases as time flows, which might be explained by habituation effects due to presenting the same visual object. On the other hand, absolute correlation and MI show more fluctuations and do not present the same increasing pattern.

---

Table 6-1. Two-sample Kolmogorov-Smirnov test results for PE and WPE. A single subject was considered in the analysis.

<table>
<thead>
<tr>
<th>KS Test</th>
<th>PE</th>
<th>WPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null hypothesis rejection</td>
<td>False</td>
<td>True</td>
</tr>
<tr>
<td>$p$-value</td>
<td>0.508</td>
<td>0.009</td>
</tr>
<tr>
<td>Test statistic</td>
<td>0.101</td>
<td>0.202</td>
</tr>
</tbody>
</table>
Figure 6-3. The averaged phase-locking values between the reference channel and all the other channels for (a) the face condition, and (b) the Gabor patch condition. The whole time series were used to calculate the PLVs, resulting in a reduced time resolution. The

Figure 6-4. The empirical cumulative distributions of the averaged phase-locking values for the face and Gabor patch conditions. A clear distinction between the two conditions can be observed. This comes however at the expense of a sacrificed time resolution.

**PE and WPE:** A precursor for a useful usage of PE (WPE) within the above context is to assign a “complexity” curve for each recorded signal, corresponding to an array of PE (WPE) values computed over a given time window (114 ms in this case). We can then compute the dependence between the different channels by simply applying correlation on these curves. Intuitively, this implies using a linear measure of dependence
to measure how close the complexity of two time series are. In our simulations, we select Spearman’s rho as a measure of statistical dependence between the different PE (WPE) curves. In Figure 6-10, the obtained correlation values are mapped onto the corresponding locations on the human scalp. In the case of WPE (Figures 6-10.A and 6-10.B ), more activity can be spotted in locations that seem to point towards sources in the occipito-parieto-temporal area of the right brain hemisphere. This outcome aligns with the results obtained in [Fadlallah et al., 2012b], which, as previously mentioned, indicate higher activity in that specific region. On the other hand, PE tends to show activity localized in right posterior areas.

**Kendall’s tau and TGMA:** Figure 6-12 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 Figure 6-13 (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
Figure 6-6. Same procedure as in Figure 6-5 applied for GMA. The upper row corresponds to forward computations of GMA, i.e. from the selected channel to the others while the lower one shows the backward computations. More dependence can be seen for the two conditions in the right hemisphere. The noticeable active region for the face condition seems to correspond to medial occipitotemporal structures and might reflect the activity of the fusiform face area. The asymmetry of the measure does not significantly affect the locations of the active regions.

can be seen in Figure 6-14. 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.
Figure 6-7. Same procedure as in the previous figure (Figure 6-6) applied on the second subject. The similarity in obtained results supports the validity of the analytic procedure across multiple individuals.

Figure 6-8. Same procedure as in Figure 6-7 applied for mutual information. Unlike correlation and GMA, mutual information shows less discriminability between the two conditions and relatively similar levels of dependence between the two brain hemispheres.
6.4.2 All Channels

6.4.2.1 Displaying the dynamic dependence graph

An interactive display of channel dependencies evolving over time is generated for each dependence measure and condition. Every dynamic dependence graph consists of 36 frames. For convenience, we present a total of 12 snapshots, each representing the average
Figure 6-11. Empirical cumulative distribution functions (CDFs) per condition for PE (left) and WPE (left). Data from one subject was used to generate this figure. PE shows indiscernable distributions of dependence, unlike WPE, where the curves are more distinguishable.

Figure 6-12. 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.

of three frames. Further extraction of graph theoretical measures from the shown graphs can be seen in the next section (Section 6.4.2.2). Figures 6-15, 6-16, 6-17 and 6-18 display dynamic dependence graphs for TGMA and absolute correlation, for the two conditions. In each figure, 12 frames with their corresponding time stamps are shown. TGMA shows more pronounced local and long-range dependencies for the face condition than for the Gabor patch condition. Absolute correlation, on the other hand, seems to have a lot of spurious connections for the two conditions.
Figure 6-13. Generating null hypotheses of uncorrelatedness for TGMA and Kendall’s tau using EEG data. The top row shows the null hypotheses for uncorrelatedness generated by randomly permuting time series samples and the distribution of the obtained p-values, computed with respect to the generated distributions. The bottom row plots the obtained p-value per channel per condition for TGMA / Kendall’s tau.

6.4.2.2 Local descriptors using graph theoretical measures

To further characterize the main network structures involved in discriminating the face and Gabor patch conditions in the dependence graphs, we compute graph theoretical quantities per node, using the difference of the face and Gabor patch dependence matrices, and then statistically processing the matrices to discard values falling within two standard deviations of the mean. Two representative measures of dependence are used when extracting graph theoretical quantities from dependence matrices computed using all the channels: the first is the time-series generalized measure of association (TGMA) [Fadlallah et al., 2012a], and the second is Spearman’s rho [Spearman, 1904], a nonparametric measure of correlation. For each measure of dependence, we compute five node quantities, namely the degree, clustering coefficient, betweenness centrality, subgraph centrality
Figure 6-14. The dashed line corresponds to the significance level of TGMA and the solid line corresponds to the significance level of Kendall’s tau. TGMA has 51 significant links for the face condition and 32 for the Gabor patch condition. For Kendall’s tau, the numbers are respectively 90 and 72, which reflects the discriminatory power of TGMA.

and local efficiency. Computations are made per windows of time corresponding to 114 ms, to achieve a better time resolution and allow tracking any time-varying activity. Alternatively, the whole time series can be used to obtain mean assessment values. The obtained values are visualized in sensor space to identify the active regions involved. Figure 6-19 shows the resulting plots (averaged over the time windows) for GMA and Figure 6-20 those for Spearman’s rho.

In Figure 6-21, we show how betweenness centrality changes with time. Each of the displayed subplots corresponds to computations extracted from 1 second of data. We can observe a consistency in the active regions across time, especially towards the later time windows, which suggests the reinforcement of communication between sources as time passes.

Figure 6-22 shows the connected components corresponding to each condition. The size of the corresponding connected component was plotted for each channel. The size of the main connected component for the “Face” condition is substantially higher than that
Figure 6-15. Dynamic dependence graph using TGMA (Face condition). Data from Subject 3 is used. The dependence matrices were averaged over trials and processed as to discard all values falling within two standard deviations of the mean. 12 frames with their corresponding times are shown. More pronounced local connectivity patterns can be observed for the face condition, with more involvement of the $C_z$, $C_4$ and $F_4$ clusters. To further characterize the main network structures involved in discriminating the face and Gabor patch conditions, graph theoretical quantities are computed per node, using the difference of the face and Gabor patch dependence matrices (Refer to Figure 6-19).

of the “Gabor patch” condition. Figure 6-23 maps the different in active regions to the sensor space.
Figure 6-16. Dynamic dependence graph using TGMA (Gabor patch condition). Data from Subject 3 is used. The dependence matrices were averaged over trials and processed as to discard all values falling within two standard deviations of the mean. 12 frames with their corresponding times are shown. More pronounced local connectivity patterns can be observed for the face condition, with more involvement of the Cz, C4 and F4 clusters. To further characterize the main network structures involved in discriminating the face and Gabor patch conditions, graph theoretical quantities are computed per node, using the difference of the face and Gabor patch dependence matrices (Refer to Figure 6-19).

6.4.2.3 Global descriptors using information theoretic concepts

Distributions of mutual information between matrices belonging to the same or different conditions were generated and are shown in Figure 5-2.
Figure 6-17. Dynamic dependence graph using absolute correlation (Face condition). Data from Subject 3 is used. The dependence matrices were averaged over trials and processed as to discard all values falling within two standard deviations of the mean. 12 frames with their corresponding times are shown. To further characterize the main network structures involved in discriminating the face and Gabor patch conditions, graph theoretical quantities are computed per node, using the difference of the face and Gabor patch dependence matrices (Refer to Figure 6-20).

6.4.3 Internal Consistency

We use Cronbach’s alpha as a measure of internal consistency. Since the recordings did not all have the same number of channels, we roughly align our data by mapping every electrode $i$ in the 257-setting to its $[i/2]$ counterpart in the 129-setting. Since such assignment of electrodes is not accurate and does not necessarily coincide with the exact locations on the scalp, the computed values of Cronbach’s alpha are approximate and can
Figure 6-18. Dynamic dependence graph using absolute correlation (Gabor patch condition). Data from Subject 3 is used. The dependence matrices were averaged over trials and processed as to discard all values falling within two standard deviations of the mean. 12 frames with their corresponding times are shown. To further characterize the main network structures involved in discriminating the face and Gabor patch conditions, graph theoretical quantities are computed per node, using the difference of the face and Gabor patch dependence matrices (Refer to Figure 6-20).

be thought of as lower bounds of consistency (this is since we hypothesize that an accurate mapping will most probably increase the observed consistency). Cronbach’s alpha was computed per condition per dependence measure.

Two sets of computations were performed. In the first, the dependence matrices were averaged over channels and time windows for each subject and condition, resulting in two $N_c \times N_t$ matrices per subject (Table 6-2). In the second, the dependence matrices were
Figure 6-19. Several measures extracted from two dependence graphs constructed using TGMA. First, we compute the difference between the two dependence matrices corresponding to the face and Gabor patch conditions (averaged over trials), and discard the values falling within 2 standard deviations of the mean. We then use graph theoretical measures to characterize the importance of each node. The measures include from left to right: (a) the node degree, (b) the node clustering coefficient, (c) the node betweenness centrality, (d) the node subgraph centrality, and (e) the node local efficiency. Computations were performed using windows of 114 samples and plots were averaged over 36 windows.

averaged over channels and time windows then concatenated from different subjects to form two $N_c \times N_t$ matrices. The results for this approach are shown in Table 6-3.

Table 6-2. Cronbach’s alpha per subject computed for different dependence measures and conditions. The dependence matrices were averaged over time windows and the consistency was computed across trials.

<table>
<thead>
<tr>
<th>Dependence measure</th>
<th>Face</th>
<th>Gabor patch</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGMA</td>
<td>0.9336</td>
<td>0.9322</td>
</tr>
<tr>
<td>Kendall’s tau</td>
<td>0.9672</td>
<td>0.9671</td>
</tr>
<tr>
<td>Spearman’s rho</td>
<td>0.9635</td>
<td>0.9634</td>
</tr>
</tbody>
</table>

6.4.4 Performance Analysis

6.4.4.1 The KS test

Figure 6-24 shows the empirical distributions per condition across the sensor network for each of the considered dependence measures and the corresponding value of the KS statistic. Distributions for absolute correlation and mutual information are more distinct
Figure 6-20. Several measures extracted from two dependence graphs constructed using Spearman’s rho. First, we compute the difference between the two dependence matrices corresponding to the face and Gabor patch conditions (averaged over trials), and discard the values falling within 2 standard deviations of the mean. We then use graph theoretical measures to characterize the importance of each node. The measures include from left to right: (a) the node degree, (b) the node clustering coefficient, (c) the node betweenness centrality, (d) the node subgraph centrality, and (e) the node local efficiency. Computations were performed using windows of 114 samples and plots were averaged over 36 windows. Again, computations were performed using windows of 114 samples and plots were averaged over 36 windows.

Figure 6-21. The betweenness centrality per node over windows of 1 sec of data. First, we compute the difference between the two TGMA dependence matrices corresponding to the face and Gabor patch conditions, and discard the values falling within 2 standard deviations of the mean. Every 9 time windows were averaged to yield a resolution of approximately 1 second. We then compute the betweenness centrality per node and project it to the scalp.
Figure 6-22. The size of the connected components per channel for the face and Gabor patch conditions. The dependence measure used is TGMA and the third subject was used. The dependence matrices were averaged over trials and time windows, then statistically processed to discard values falling within two standard deviations of the mean. The connected components were then extracted from the resulting adjacency matrix and the size of the connected component at each node is displayed.

than correlation but show less discriminability than GMA. Looking at a time window of 114 samples, the test statistic obtained for GMA is 0.9125, compared to 0.3441 and 0.4844 for mutual information and absolute correlation, as shown in Figure 6-25. All measures performed better than the power classification scheme in Figure 2-12. Results were similar when using a longer time window with a 15% difference for GMA, when comparing to a smaller time window. The range of values drops for all measures of dependence measures but this comes at the expense of reduced time resolution. Table 6-4 shows the KS test results obtained per subject.
Figure 6-23. The cardinality of the difference in connected components sizes shown in Figure 6-22 mapped to sensor space. Each node displays the difference between the sizes of the largest connected components including the node, pertaining to each condition.

Table 6-3. Cronbach’s alpha across subjects computed for different dependence measures and conditions. The dependence matrices were averaged over time windows and concatenated from all participants.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Condition</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Subj 1</td>
</tr>
<tr>
<td>TGMA</td>
<td>Face</td>
<td>0.9375</td>
</tr>
<tr>
<td></td>
<td>Gabor</td>
<td>0.9221</td>
</tr>
<tr>
<td>Kendall’s tau</td>
<td>Face</td>
<td>0.9205</td>
</tr>
<tr>
<td></td>
<td>Gabor</td>
<td>0.9032</td>
</tr>
<tr>
<td>Spearman’s rho</td>
<td>Face</td>
<td>0.9034</td>
</tr>
<tr>
<td></td>
<td>Gabor</td>
<td>0.8777</td>
</tr>
</tbody>
</table>

6.4.4.2 Classification results

We train with data gathered from 5 participants to generate distributions of entropy estimates for the two conditions, then use the 6th participant for testing. Therefore we have 100 training matrix samples per condition, or 200 in total. The approach is supervised in that the labels for the training matrices are known. The 20 trials belonging
Figure 6-24. Empirical cumulative distribution functions (CDFs) computed from dependence values across the 129 electrodes and averaged over trials for: (a) Pearson’s correlation, (b) Absolute Pearson’s correlation, (c) mutual information and (d) GMA. Time window used is 114 samples or 114 ms. The error bars show the variability with time windows (38 windows in this case).

Figure 6-25. Test statistic for the two-sample KS test applied for each of the four dependence measures. Points in red were obtained when using time windows of 114 samples and those in blue when using 456 samples. The dotted green line represents the KS statistic obtained using power distributions.
Table 6-4. KS test results for all subjects. The dependence matrices were averaged over trials and time windows.

<table>
<thead>
<tr>
<th>Measure</th>
<th>KS test</th>
<th>Subjects</th>
<th>Subj 1</th>
<th>Subj 2</th>
<th>Subj 3</th>
<th>Subj 4</th>
<th>Subj 5</th>
<th>Subj 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>TGMA</td>
<td>KS result</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS p-value</td>
<td>0.0487</td>
<td>0.0058</td>
<td>0.0613</td>
<td>0.0047</td>
<td>0.8302</td>
<td>0.0047</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS statistic</td>
<td>0.3973</td>
<td>0.2093</td>
<td>0.1775</td>
<td>0.1518</td>
<td>0.1545</td>
<td>0.1518</td>
<td></td>
</tr>
<tr>
<td>Kendall’s tau</td>
<td>KS result</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS p-value</td>
<td>0.0001</td>
<td>0.6112</td>
<td>0.5075</td>
<td>0.0013</td>
<td>0.9886</td>
<td>0.0156</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS statistic</td>
<td>0.4264</td>
<td>0.0932</td>
<td>0.1008</td>
<td>0.1673</td>
<td>0.0389</td>
<td>0.1362</td>
<td></td>
</tr>
<tr>
<td>Spearman’s rho</td>
<td>KS result</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS p-value</td>
<td>0.0001</td>
<td>0.7175</td>
<td>0.5075</td>
<td>0.0013</td>
<td>0.9696</td>
<td>0.0044</td>
<td></td>
</tr>
<tr>
<td></td>
<td>KS statistic</td>
<td>0.4651</td>
<td>0.0853</td>
<td>0.1008</td>
<td>0.1673</td>
<td>0.0428</td>
<td>0.1556</td>
<td></td>
</tr>
</tbody>
</table>

The classification results using this approach are shown in Table 6-5.

Table 6-5. Cronbach’s alpha across subjects computed for different dependence measures and conditions. The dependence matrices were averaged over time windows and concatenated from all participants.

<table>
<thead>
<tr>
<th>Classification rate for Subject 6 per condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependence measure</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>TGMA</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Kendall’s tau</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Spearman’s rho</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
6.5 Discussion

Results show active regions in the occipito-parietal part of the brain for both conditions with a greater dependency between occipital and inferotemporal sites for the face stimulus. This aligns with previous evidence suggesting re-entrant organization of the ventral visual system, showing heightened re-entry when viewing meaningful or salient stimuli. The pattern of active regions could be traced across the different measures of dependence with a better classification achieved by GMA/TGMA. Moreover, the slower variation of the dependence level as time passes seems to point towards habituation effects due to presenting the same visual object. Classification using two time windows (114 and 456 samples) performed better than the simple power classifier as shown in Figure 6-25. Moreover, a good internal consistency was observed since all subjects had values of Cronbach’s alpha near or above 0.9.
CHAPTER 7
CONCLUSIONS AND FUTURE DIRECTIONS

7.1 Summary and Discussion

In this dissertation, we examined ssVEPs evoked by flickering stimuli showing a human face and Gabor patches. Dependencies between channel locations were computed in time windows of 114 ms for several traditional and novel measures of dependence.

A suitable preprocessing framework was designed to analyze dependence in current CSD time series extracted from recorded EEG. The framework is robust in that desired output achieves minimal sensitivity with respect to free parameters. The main parameter to be considered when filtering is the quality factor. Graph and information theoretical frameworks were designed to explore the discriminability between conditions. The performance was studied in terms of discriminating between the two conditions. Results show consistency in the active brain regions over cycles. More activity is visible in the right hemisphere of the brain for the Face condition, near $P_4$. A high level of dependence is noticed between the occipital and right parieto-temporo-occipital sensor locations. This might be explained by a communication between sources in the primary visual cortex and ventral region sources. This is also consistent with engagement of the fusiform or right occipito-temporal areas. Proposed measures of association showed more discriminability for the two stimuli than standard measures.

Estimates of source locations underlying the ssVEP activity strongly suggest that most scalp-recorded ssVEP signals originate from lower-tier visual cortex. This is inline with the findings by Russo et al. [2007] and Wieser and Keil [2011] and those in Muller et al. [1997] that localize the visual ssVEP in posterior occipital and ventral occipital cortex. Russo et al. [2007] further suggest that simple ssVEP models consisting of one or two sources in the occipital region of the brain are unable to explain the ssVEPs magnitude and phase, which favors the hypothesis of a distributed source system with enhanced communication between the occipital/parieto-occipital region and sources in
ventral regions. Potential sources suggested by this study concur with sources of face-specific responses found by Sams et al. [1997].

7.2 Future Work

Future work would proceed in one or more of the following directions:

1. First we propose to incorporate to our analysis measures of signal regularity or complexity. This can be done according to two ways:
   (a) Adopting a sliding window analysis using measures of signal complexity. Such approach can yield descriptors of data that keep temporal dependence to a minimum, which facilitates the usage of measures of dependence at a second stage.
   (b) Deriving novel measures of association using correntropy and entropy estimators. Such approach would use again measures of complexity, combined either with nearest neighbors or ordinal ranks, in order to quantify directly dependence between two variables.

2. Improving the quality of the statistical tests used. This is since the KS test is sensitive to deviations in the empirical CDF functions of the two samples being compared. We can either choose to set confidence intervals of the KS test using a bootstrapping approach or go beyond the KS test to generate the null hypothesis of distribution equality by using a resampling approach. The **bootstrap approach** would proceed according to the following steps:
   (a) Drawing a resample of a certain size (say half the number of trials) with replacement from the first sample and a separate resample of the same size from the second sample.
   (b) Choosing a statistic that compares the two resamples (can be KS or Mann-Whitney).
   (c) Repeating this resampling process hundreds of times.
   (d) Constructing the bootstrap distribution of the statistic.
3. Adapting the methodology for use with different datasets. Examples include presenting different pictures with varying emotional context and checking the ability to discriminate between the neutral and more threatening or emotionally involving pictures.

Although we did not study directional connectivities in this context, the location of active regions witnessed for the face stimulus is consistent with findings suggesting re-entrant modulation of early visual cortex, originating from higher tiers and entering lower tiers of visual cortex [Keil et al., 2009]. On the other hand, to further elaborate on the obtained results, it would be interesting to infer more information about the nature of communication between the occipito-parietal cortex and the inferotemporal and whether it involves intermediate sources. This can be done by studying the time delay properties of the obtained signals over a set of channels. Another graph theoretical approach would be to work directly on the adjacency matrix via graph matching procedures, using a procedure that automatically matches two graphs displaying dependence structure. Previous examples in the literature where graph comparison and matching methods were applied include the papers by Conte et al. [2004] and Zegura et al. [1997].
Current source density or CSD measures in a reference-free fashion the strength of extracellular current generators underlying the recorded raw EEG. The computation of CSD is based on a linear volume conduction model [Nicholson, 1973; Nicholson and Freeman, 1975]. To derive the CSD measure, we start from the very well known Ohm’s law:

\[ \mathbf{J} = \sigma \mathbf{E} \]  (A–1)

where \( \mathbf{J} \) denotes the current flow density, \( \mathbf{E} \) the corresponding electric field, and \( \sigma \) the conductivity of the medium. Note that \( \mathbf{E} \) can be written as:

\[ \mathbf{E} = -\nabla V \]  (A–2)

where \( V \) denotes the scalar value of the electric potential (or voltage) and \( \nabla \) denotes the gradient operator. Using Equation A–1 and A–5 and applying the divergence operator on both sides of the equation.

\[ I_d = \nabla \cdot \mathbf{J} = \nabla \cdot \sigma (-\nabla V) \]  (A–3)

Hence:

\[ I_d = -\sigma \nabla \cdot (-\nabla V) = -\sigma \nabla^2 V = -\sigma \Delta V \]  (A–4)
Thus CSD is proportional to the second spatial derivative (or Laplacian) of the measured field potential [Tenke and Kayser, 2005]. In Cartesian coordinates, the Laplace operator can be defined as:

\[
\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}
\]  

(A-5)

The Surface Laplacian (SL) can be then estimated using two different approaches, a local one and a global one. The main local approach was proposed by Hjorth [1975], in which the SL is estimated directly at selected sites. The method simply proceeds by computing the difference between the potential at each electrode site and the average potential of the nearest four neighbors. Such method assume that distances separating electrodes are equal, just like the angles built by the electrodes configuration [Tandonnet et al., 2005].

On the other hand, the global approach uses all electrodes to compute the SL. It proceeds by constructing an interpolation function to represent the potential distributions and estimating the SL based on that function. Most global techniques are based on spline interpolation methods as done by Nunez and Cutillo [1995], Perrin et al. [1989] and Babiloni et al. [2001].

A.2 Computing the CSD

A.2.1 Forward and Inverse Problems

Let \( P(r, t) \) be the electric dipole moment per unit volume. Scalp potential may be expressed as a volume integral of \( P(r, t) \) over the entire brain provided \( P(r, t) \) is defined generally rather than in columnar terms [Srinivasan and Nunez, 2007]. For the important case of dominant cortical sources, scalp potential may be approximated by the following integral over the cortical volume \( \phi \), where \( G(r, rt) \) refers to a Green’s function containing
all geometric and conductive information about the head volume conductor and weights
the integral accordingly [Srinivasan and Nunez, 2007].

\[ V_S(r, t) = \iiint G(r, rt) \cdot P(rt, t) d\Theta(rt) \]  
(A-6)

It might be useful here to remind that the forward problem in EEG consists in
tuning the head model and then using the corresponding Green’s function, perform the
calculations in Equation A–6 for an assumed source distribution. On the other hand, the
inverse problem consists of starting with the recorded scalp potential distribution \(V_S(r, t)\)
under some constraints on \(P(r, t)\) in order to retrieve the best source distribution \(P(r, t)\).
The inverse problem has no unique solution, i.e. for any given distribution of the electrical
potential on the surface of the head, there exists an infinite number of possible source
configurations inside the head [Helmholtz, 1858; Junghofer et al., 1997].

A.2.2 CSD and Cortical Mapping

The computation of the cortical mapping and the current source density (CSD)
can be obtained through the same procedure when using spherical spline interpolation.
 Modeling the volume conductor as a spherically-shaped isotropic volume consisting of four
layers (scalp, skull, cerebro-spinal fluid or CSF and brain), different conductivities \(\sigma_i\) are
assigned to each layer. Radial sources are assumed in the derivation. Cuffin and Cohen
[1979] express the potential caused by a single radial current dipole, located on the surface
of the sphere as:

\[ V(\theta, \phi) = p \sum_{n=1}^{\infty} P_n^\theta(\cos \theta) \cdot \Gamma^{scalp}(n) \]  
(A-7)

where \(\Gamma^{scalp}(n)\) is given by:

\[ \Gamma^{scalp}(n) = \frac{1}{4\pi \sigma_4 R^2} \times \frac{(2n+1)^4.f^{n-1}.(cd)^{2n+1}}{\Gamma(n)} \]  
(A-8)
where in the above equations:

- $b$, $c$ and $d$ describe a proportional relation to the corresponding shell radius.
- $\Gamma(n) = f(b, c, d, \sigma_i)$.
- $\theta$ and $\phi$ denote the polar and azimuth angles in spherical coordinates.
- $P_n$ denotes a Legendre polynomial of order $n$.

The solution can be cast as an eigenvalue problem when written in matrix notation:

$$V = p G_{ij}^{scalp}$$

(A-9)

where $G_{ij}^{scalp} = \sum_{n=1}^{\infty} P_n. (\cos \alpha_{ij}). \Gamma^{scalp}(n)$ and $\alpha_{ij}$ refers to the angle formed by the pair of electrodes $i$ and $j$ ($i, j \in 1, ..., N$).

It is important to note that casting the problem in the above format is useful and efficient. This is since the term $G_{ij}^{scalp}$ only depends on the model parameters and is independent of the measurements, i.e. it can be computed and stored then applied to different measurement matrices.

At this stage, the idea is to interpolate the scalp potential and use the result to compute the CSD. This can be done using an important property of the Laplacian in spherical coordinates, i.e.

$$\Delta_{\theta,\phi} Y_l^m(\theta, \phi) = -l(l + 1)Y_l^m(\theta, \phi)$$

(A-10)

that results in having the spherical Laplacian of a Legendre polynomial equal to an integer multiple of the same function, or $\Delta P_n = -n(n + 1)P_n$. Hence we can write $C^{scalp}$ as

$$C^{scalp} = -\sum_{n=1}^{\infty} P_n(\cos \alpha_{ik}). n.(n + 1). \Gamma^{scalp}(n)$.

A similar expression is derived for the CSD of the cortical potential distribution.
A figure illustrating the potential versus CSD maps can be seen in Figure A-1.

Figure A-1. **Left:** Potential map versus **right:** CSD map for a visual evoked potential (VEP) recording. More activity localization can be seen in the CSD map.

---

1 The figure has been reprinted from [Van de Wassenberg, 2008].
A least-squares linear-phase filter design is a simple and optimal (with respect to the square error criterion) method to design FIR filters.

Let $R(f)$ and $D(f)$ denote the obtained and desired amplitude responses and $\Gamma(f)$ be a positive weighting function whose values change according to the pass, transition and stop bands. Our goal is to find $h(n)$ such that $e = \int_{f=0}^{f=1/2} \Gamma(f).[R(f) - D(f)]^2 df$ is minimized. Assuming a Type I filter, we have:

$$R(f) = \sum_{n=0}^{M} r(n).\cos(2\pi nf)$$  \hspace{1cm} (B-1)

where in the above equation the relation between $h(n)$ and $r(n)$ is known. The coefficients of $r(n)$ can be hence retrieved by taking the derivative of the integral defining $e$. This would give:

$$\sum_{n=0}^{M} r(n) \int_{f=0}^{f=1/2} \Gamma(f) \cos(2\pi nf) \cos(2\pi kf) df = \int_{f=0}^{f=1/2} \Gamma(f) D(f) \cos(2\pi kf) df$$  \hspace{1cm} (B-2)

This can be written as $\Phi r = \beta$ where:

$$\beta(k) = \frac{1}{\pi} \int_{f=0}^{f=1/2} \Gamma(f) D(f) \cos(2\pi kf) df$$  \hspace{1cm} (B-3)

and:

$$\Phi(k, n) = \frac{1}{\pi} \int_{f=0}^{f=1/2} \Gamma(f) \cos(2\pi nf) \cos(2\pi kf) df$$  \hspace{1cm} (B-4)

Hence finally, $r(n)$ can be obtained as $r(n) = \Phi^{-1} \beta$. 

133
APPENDIX C
SECOND-ORDER BUTTERWORTH FILTERS

The magnitude response of second-order Butterworth filter can be expressed as:

\[ |H(jw)| = \frac{1}{\sqrt{1 + w^4}} \quad (C-1) \]

We can then obtain the minimum phase transfer function from the magnitude response as:

\[ H(s) = \frac{1}{1 + \sqrt{2}s + s^2} \quad (C-2) \]

The corresponding amplitude response function can be then deduced as:

\[ H(jw) = \frac{1}{(1 - w^2) + j(\sqrt{2}w)} \quad (C-3) \]

The phase response would then be:

\[ \phi(w) = -\tan^{-1}\frac{\sqrt{2}w}{1 - w^2} \quad (C-4) \]

And the group delay:

\[ \tau(w) = \frac{\sqrt{2}(1 + w^2)}{1 + w^4} \quad (C-5) \]
APPENDIX D
WEIGHTED-PERMUATION ENTROPY

An adequate testing scheme to weighted-permutation entropy (WPE) would include
spiky data because it poses a challenge to a simple motif count approach and exhibits
sudden changes. Simulations were performed on both synthetic data and EEG data.

D.1 Synthetic Data

As a first motivation, we suggest to analyze the behavior of PE and WPE in presence
of an impulsive and noisy signal. Figure D-1.a shows 1000 samples of a signal consisting of
an impulse and additive white Gaussian noise (AWGN) with zero mean and unit variance.
Windows of 80 samples slid by 10 samples were used and results were averaged over 10
simulations. A remarkable drop in the value of WPE is noticed in the impulse region. No
marked change can be observed in the case of PE for the same region.

As next step, we try a train of Gaussian-modulated sinusoidal pulses with decaying
amplitudes. The value of $\tau$ was set to 1. Sliding windows of 50 samples with increments
of 10 samples were used and $m$ was set to 3. Again, the signal was corrupted by AWGN
and simulations were run across different variance levels. Figure D-2 shows the variations
of the signal’s entropy for four different methods. The performance of PE and WPE is
compared to two other methods from the literature, namely approximate entropy or ApEn
[Chon et al., 2009; Pincus, 1991] and the composite PE index or CPEI [Olofsen et al.,
2008]. In the following, we give a brief description of each.

Approximate Entropy (ApEn or AE): Approximate entropy is a measure that
quantifies the regularity or predictability of a time series. It is defined with respect to a
free parameter $r$ as follows:

$$H_a = \Phi^m(r) - \Phi^{m+1}(r)$$ (D–1)
Figure D-1. (Color online) PE versus WPE in the case of an impulse. (a) Impulse with additive white Gaussian noise with zero mean and unit variance. (b) Computed PE and WPE values with windows of 80 samples slid by 10 samples. A remarkable drop in the value of WPE is noticed in the impulse region for which PE values do not show any marked change.

where $\Phi^m(r)$ is defined as:

$$\Phi^m(r) = \frac{1}{N - (m - 1)\tau} \sum_{i=1}^{N-(m-1)\tau} \ln C^m_i(r)$$  \hspace{1cm} (D-2)

and $C^m_i(r)$ is defined using the Heavyside function $\Theta(u)(1$ for $u > 0, 0$ otherwise) and a distance measure $\text{dist}$:

$$C^m_i(r) = \frac{\sum_{j=1}^{N-(m-1)\tau} \Theta(r - \text{dist}(X_m^{m,\tau}, X_j^{m,\tau}))}{N - (m - 1)\tau}$$  \hspace{1cm} (D-3)
Here the value of $r$ is set to be 0.2 times the data standard deviation as per the thorough discussion in [Chon et al., 2009]. The distance measure we use is the same suggested in [Pincus, 1991] and can be formulated as:

$$\text{dist}(x_{i}^{m,\tau}, x_{j}^{m,\tau}) = \max_{k=1,\ldots,m} \left| x_{i+(k-1)\tau} - x_{j+(k-1)\tau} \right|$$

**Composite PE index (CPEI):** The composite PE index (CPEI) is an alteration of permutation entropy that differentiates between the types of patterns. It is calculated as the sum of two permutation entropies corresponding to motifs having different delays where the latter (denoted as $\tau$ in this dissertation) is determined by whether the motif is monotonically decreasing or increasing. CPEI, which we denote by $H_i$ in this dissertation, responds rapidly to changes in EEG patterns and can be defined as follows [Olofsen et al., 2008]:

$$H_i = \frac{1}{\ln(m! + 1)} \frac{H(m, 1) + H(m, 2)}{2}$$  \hspace{1cm} (D-4)

The normalization denominator in Equation D–4 consists of the original number of motifs in addition to a newly introduced motif to account for ties (ties describe cases where negligible differences in amplitude occur within a motif). As a side note, the averaging step performed in that equation is highly approximative because of the lack of independency between motifs at different delays.

It is noticeable that WPE consistently drops for portions of the signal showing pulses. This is desired because of the lesser complexity of these regions and expected because of their immunity to noise. Here we assume that the information contained in the examined signals is amplitude-dependent. Such results meet our expectations since WPE is clearly able to differentiate between bursty and stagnant regions of the pulse train. In other words, using the variance contributes to weakening the noise effects and assigning more
Figure D-2. (Color online) Different entropy measures (PE, WPE, CPEI and AE) applied on a Gaussian-modulated sinusoidal train with a frequency of 10 kHz, a pulse repetition frequency of 1 kHz and an amplitude attenuation rate of 0.9. Initial signal was corrupted by additive white Gaussian noise (AWGN) having mean $\mu = 0$ and variance $\sigma^2 = 0.2$. The sampling rate was 50 kHz and computations used a 50-sample sliding window with increments of 10 samples. The recorded SNR was of 4.8 dB.
weight to the regular spiky patterns corresponding to a higher amount of information, which results in easier predictability and less complexity. It is important to note two things: (1) the contribution of patterns with higher variance towards the value of WPE dominates those of patterns with lesser variance which highlights the powerfulness of the method in detecting abrupt changes in the input signal and (2) the fact that WPE is computed within a specific time window explains why WPE values corresponding to impulsive segments of the signal do not decrease in spite of the decreasing amplitudes of the spikes (the normalization effect in (4) takes place within each window). We also plot in Figure D-3 the values of PE and WPE for different levels of signal-to-noise ratio (SNR). As anticipated, both entropy measures decrease with the increase of the SNR since the effect of noise contributing to more complexity becomes less significant. WPE decreases at a higher pace than PE, which reflects a better robustness to noise. As a final note on this section, we point out that traditional methods like zero-crossing spike detection techniques might be useful for the purpose of this simulation, however the sought goal was to demonstrate, using synthetic data the ability of WPE to discriminate between regimes of data.

![Figure D-3](image-url)

Figure D-3. (Color online) Normalized PE and WPE values for different SNR levels. The signal used is the same as in Figure D-2.
D.2 Single-Channel EEG Data Analysis

In Figure D-4, the same comparisons are performed for a sample EEG recording processed as in [Fadlallah et al., 2011]. Highpass filtering was further applied on the signal because we are interested in removing very low frequency components. It can be seen that WPE locates the regions where abrupt changes occur in the initial signal more accurately than the other methods, which is inline with our original expectations. The same is reflected in Figure D-5 that shows a processed EEG portion corresponding to another channel. Our results show that increasing $m$ beyond 4 affects the running time without significantly changing the obtained entropies. This is inline with the findings in [Lehnertz, 2007] where the parameter selection problem has been addressed and [Olofsen et al., 2008]. For situations where the effect of $m$ is more pronounced, the running time issue can be addressed by speeding up the sliding of the window as this entails a higher number of affected patterns at each instance.

D.3 Epilepsy Detection

Setting: Next we propose to apply WPE for epilepsy detection. We use the same data as Quiroga et al. [1997; 2002], in which tonic-clonic seizures of a subject were recorded with a scalp right central electrode (located near C4 in a standard 10 – 20 montage). The recording consisted of 3 minutes, including around 1 minute of pre-seizure time and 20 seconds of post-seizure activity. A sampling rate of 102.4 Hz was used to collect the signal.

Discussion: We computed different measures of entropy on windows of 50 samples of data slid by 5 samples (Figure D-6.b). The obtained curves are further smoothed in Figure D-6.c using a moving average filter of length 35 samples. The commencement of epileptic activity in the recorded signal induces noticeable changes for all entropy measures, in
Figure D-4. (Color online) Analysis performed on filtered EEG data sampled at 1000 Hz and processed as in [Fadlallah et al., 2011]. WPE outperforms other entropy measures in location regiments exhibiting abrupt changes in the signal. The window length used for this plot was 114 with an overlap of 2 samples.

particular for WPE that exhibits a significant jump in value. This is further quantified by computing the ratio of average measured entropies of epileptic and non-epileptic segments (Table D-1), which shows a more pronounced difference between both portions for WPE. The latter achieves almost twice better discriminability between the two portions of the signal, i.e. 42% better than the next closest measure (CPEI).
Figure D-5. (Color online) Same procedure applied on a processed EEG portion corresponding to another channel. WPE mirrors best the sharp change in the signal noticeable before $t = 850$ ms. The window size used was 200 with an overlap of 2 samples at each iteration.

Table D-1. Ratio of average measured entropy between epileptic and non-epileptic segments.

<table>
<thead>
<tr>
<th>Measure</th>
<th>Ratio$^1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PE</td>
<td>1.27</td>
</tr>
<tr>
<td>WPE</td>
<td>1.85</td>
</tr>
<tr>
<td>CPEI</td>
<td>1.30</td>
</tr>
<tr>
<td>AE</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Figure D-6. (Color online) Different entropy-based measures applied on epileptic EEG. (a) EEG recording of an epileptic subject. The recording, sampled at 102.4 Hz contains approximately one minute of pre-seizure activity and 20 seconds of post-seizure activity. (b) Different measures of entropy computed using a sliding window of 50 samples with 5 samples overlap. (c) Smoothed entropy measures curves obtained by applying a moving average filter of length 35 samples.
APPENDIX E
INFINITELY DIVISIBLE FUNCTIONS

The following section is an extract from [Sanchez-Giraldo and Principe, 2013] that contains more elaboration about infinitely divisible functions.

E.1 Negative Definite Functions and Hilbertian Metrics

Let $\mathcal{M} = (\mathcal{X}, d)$ be a separable metric space. A necessary and sufficient condition for $\mathcal{M}$ to be embeddable in a Hilbert space $\mathcal{H}$ is that for any set $\{x_i\} \subset \mathcal{X}$ of $n + 1$ points,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_j (d^2(x_0, x_i) + d^2(x_0, x_j) - d^2(x_i, x_j)) \geq 0, \text{ for any } \alpha \in \mathbb{R}^n.$$ 

This condition is equivalent to $\sum_{i,j=0}^{n} \alpha_i \alpha_j d^2(x_i, x_j) \leq 0, \text{ for any } \alpha \in \mathbb{R}^{n+1}, \text{ such that } \sum_{i=0}^{n} \alpha_i = 0.$ This condition is known as negative definiteness. Interestingly, the above condition implies that $\exp(-rd^2(x_i, x_j))$ is positive definite in $\mathcal{X}$ for all $r > 0.$ Indeed, matrices derived from functions satisfying the above property conform a special class of matrices know as infinitely divisible.

E.2 Infinitely Divisible Matrices

According to the Schur product theorem $A \succeq 0$ implies $A \odot A \odot \cdots \odot A \succeq 0$ for any positive integer $n.$ Does the above hold if we to take fractional powers of $A?$ In other words, is the matrix $A^{1/m} \succeq 0$ for any positive integer $m?$ This question leads to the concept of infinitely divisible matrices. A nonnegative matrix $A$ is said to be infinitely divisible if $A^r \succeq 0$ for every nonnegative $r.$ Infinitely divisible matrices are intimately related to negative definiteness as we can see from the following proposition

**Proposition E.2.1.** If $A$ is infinitely divisible, then the matrix $B_{ij} = -\log A_{ij}$ is negative definite.

From this fact it is possible to relate infinitely divisible matrices with isometric embedding into Hilbert spaces. If we construct the matrix:

$$D_{ij} = B_{ij} - \frac{1}{2}(B_{ii} + B_{jj}),$$

(E–1)
using the matrix $B$ from proposition E.2.1. There exists a Hilbert space $\mathcal{H}$ and a mapping $\phi$ such that:

$$D_{ij} = \|\phi(i) - \phi(j)\|^2_{\mathcal{H}}.$$  (E-2)

Moreover, notice that if $A$ is positive definite $-A$ is negative definite and $\exp A_{ij}$ is infinitely divisible. In a similar way, we can construct the matrix:

$$D_{ij} = -A_{ij} + \frac{1}{2}(A_{ii} + A_{jj}),$$  (E-3)

with the same property (E–2). This relation between (E–1) and (E–3) suggests a normalization of infinitely divisible matrices with non-zero diagonal elements that can be formalized in the following theorem.

**Theorem E.2.1.** Let $\mathcal{X}$ be a nonempty set, and let $d_1$ and $d_2$ be two metrics on it, such that for any set $\{x_i\}_{i=1}^n$, \[ \sum_{i,j=1}^n \alpha_i \alpha_j d_\ell^2(x_i, x_j) \leq 0, \] for any $\alpha \in \mathbb{R}^n$, and $\sum_{i=1}^n \alpha_i = 0$, is true for $\ell = 1, 2$. Consider the matrices $A_{ij}^{(\ell)} = \exp -d_\ell^2(x_i, x_j)$ and their normalizations $\hat{A}^{(\ell)}$, defined as:

$$\hat{A}^{(\ell)}_{ij} = \frac{A^{(\ell)}_{ij}}{\sqrt{A^{(\ell)}_{ii} A^{(\ell)}_{jj}}}.$$  (E-4)

Then, if $\hat{A}^{(1)} = \hat{A}^{(2)}$ for any finite set $\{x_i\}_{i=1}^n \subseteq \mathcal{X}$, there exist isometrically isomorphic Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, that contain the Hilbert space embeddings of the metric spaces $(\mathcal{X}, d_\ell)$, $\ell = 1, 2$. Moreover, $\hat{A}^{(\ell)}$ are infinitely divisible.

A beautiful observation from Theorem E.2.1 is that the proposed normalization procedure for infinitely divisible matrices can be thought of as finding the maximum entropy matrix among all matrices for which the Hilbert space embeddings are isometrically isomorphic.
REFERENCES


147


Kirsch, P., Besthorn, C., Klein, S., Rindfleisch, J., and Olbrich, R. (2000). The dimensional complexity of the eeg during cognitive tasks reflects the impaired information


BIOGRAPHICAL SKETCH

Bilal Fadlallah received the bachelor’s degree (BE) in computer and communications engineering and the master’s degree (ME) in electrical and computer engineering from the American University of Beirut, Lebanon, in 2006 and 2008 respectively. He joined the Computational NeuroEngineering Laboratory at the University of Florida in 2010, working on inferring functional dependencies in the human brain from EEG data. His Ph.D. work mostly lies at the intersection of several fields, including signal processing, cognitive and computational neuroscience and machine learning.

Bilal has four years of industry experience in designing and testing systems software, at Murex Systems where he lead the Murex Limits Controller team, and Microsoft where he contributed to the Windows Fundamentals team. He is the recipient of the University of Florida Certificate of Outstanding Academic Achievement and a fellow of the Lebanese National Council for Scientific Research.