A Spectral Method for Generating Surrogate Graph Signals

E. Pirondini, A. Vybornova, M. Coscia, and D. Van De Ville, *Senior Member, IEEE*

***Abstract*—The increasing availability of network data is leading to a growing interest in processing of signals on graphs. One notable tool for extending conventional signal-processing operations to networks is the graph Fourier transform that can be obtained as the eigendecomposition of the graph Laplacian. In this letter, we used the graph Fourier transform to define a new method for generating surrogate graph signals. The approach is based on sign-randomization of the graph Fourier coefficients and, therefore, the correlation structure of the surrogate graph signals (i.e., smoothness on the graph topology) is imposed by the measured data. The proposed method of surrogate data generation can be widely applied for nonparametric statistical hypothesis testing. Here, we showed a proof-of-concept with a high-density electroencephalography dataset.**

***Index Terms*—Electroencephalography (EEG), graph Laplacian, graph signals, nonparametric hypothesis testing, phase randomization, surrogate data.**

I. Introduction

NETWORK modeling and analysis is an increasingly important topic in many disciplines of science. The signal processing community has found interest in developing and tailoring classical operations to graphs [1], [2]. In this Letter, we propose to extend the widespread method of phase randomization to graph signals.

Statistical hypothesis testing works by invalidating a given null hypothesis that expresses that the measured effect is not present; e.g., it can plausibly be explained by "randomness." The method of phase randomization has been proposed to generate surrogate data under the null hypothesis that the measured signal is part of a class of stationary signals with prescribed second-order statistics [3], [4]. Practically, data (e.g., a time course or an image) are transformed into the Fourier domain where the phases of the Fourier coefficients are randomly permuted or generated. Each realization of the surrogate data is then obtained by a different randomization followed by the inverse Fourier transform (IFT). This method is widely deployed because it leads to a more realistic (and stronger) null hypothesis; e.g., direct permutation of samples would completely destroy correlation structure, while it is preserved in phase-randomized surrogate data.

Complex datasets can benefit from representations based on graphs (i.e., a topology described by vertices and edges) and graph signals (i.e., a signal that is expressed on the vertices). The emerging field of signal processing on graphs has extended and generalized already many classical signal operations to graphs such as Fourier [5], [6] and wavelet analysis [7]. In these tools, the graph Laplacian operator is used to define a spectral decomposition that is similar to the Fourier transform (FT) on regular grids [1], [8].

Here, we propose a novel method for generating surrogate graph signals that preserves correlation structure as captured by the graph Laplacian [9]. In particular, we combine the graph spectral decomposition with a sign-randomization of the graph Fourier coefficients.

Outline of the paper: Section II briefly reviews graph theory and the graph Fourier transform (GFT); Section III describes the proposed method for generating surrogate graph signals; Section IV shows the feasibility of the approach by an example graph and graph signal; Section V concludes the paper and provides extensions and future research directions.

II. Spectral Graph Transform

A. Graph Signal

Let us consider a graph $G = (X, E, \boldsymbol{W})$, where X denotes the ensemble of N vertices (i.e., $|X| = N$), E the ensemble of edges connecting the vertices, and $\boldsymbol{W}$ the adjacency matrix. The positive real-valued edge weight $W_{i,j}$ represents the strength of the connection between the ith and the jth node. In a binary graph, $W_{i,j}$ is either 0 or 1 indicating the presence of an edge; in a weighted graph, $W_{i,j}$ indicates the (positive) strength. For an undirected graph, which is the case we will assume further on, the matrix $\boldsymbol{W}$ is symmetric by construction.

A graph signal is an N-dimensional vector $\boldsymbol{x}$ that associates a value to each vertex.

B. Graph Fourier Transform

The graph Laplacian operator is defined as $\mathcal{L} = \boldsymbol{D} - \boldsymbol{W}$, where the degree matrix $\boldsymbol{D}$ is a diagonal matrix whose ith element D_i is the degree of the ith vertex; i.e., the sum of all weights of edges of this vertex. The graph Laplacian of an undirected

Manuscript received March 17, 2016; revised June 10, 2016; accepted July 16, 2016. Date of publication July 26, 2016; date of current version August 10, 2016. E. Pirondini and A. Vybornova contributed equally to this work. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Peter K. Willett.

E. Pirondini and M. Coscia are with the Bertarelli Foundation Chair in Translational Neuroengineering, Center for Neuroprosthetics École Polytechnique Fédérale de Lausanne, Lausanne1015, Switzerland (e-mail: elvira.pirondini@epfl.ch; martina.coscia@epfl.ch).

A. Vybornova is with École Polytechnique Fédérale de Lausanne, Lausanne1015, Switzerland (e-mail: anna.vybornova@epfl.ch).

D. V. D. Ville is with the Institute of Bioengineering, Center for Neuroprosthetics, École Polytechnique Fédérale de Lausanne, Lausanne1015, Switzerland, and also with the Department of Radiology and Medical Informatics, University of Geneva, Geneva 1205, Switzerland (e-mail: dimitri.vandeville@epfl.ch).

Color versions of one or more of the figures in this letter are available online at http://ieeexplore.ieee.org.

The MATLAB code used in this work can be found at http://miplab.epfl.ch/index.php/software/graph-surrogates

Digital Object Identifier 10.1109/LSP.2016.2594072

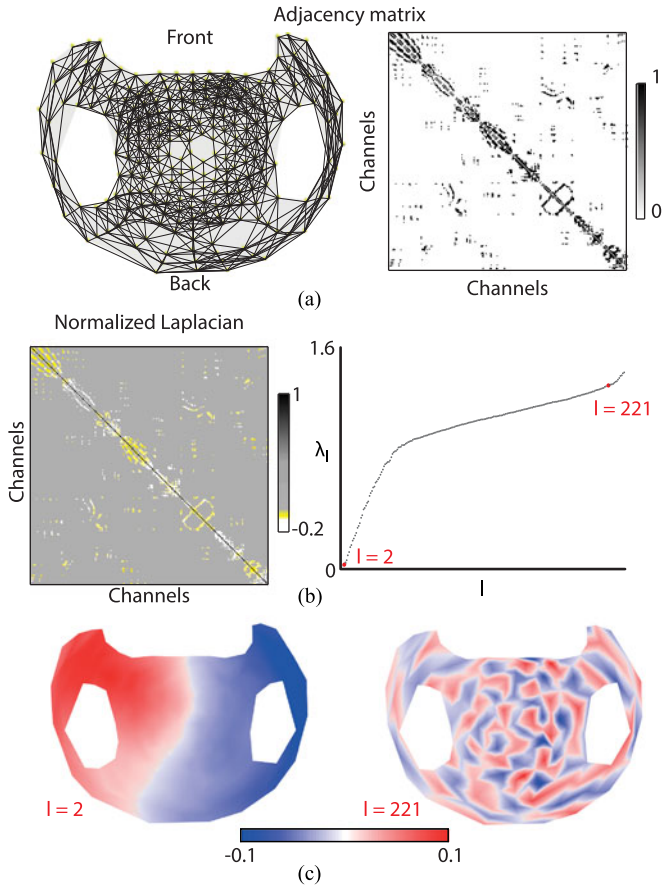

Fig. 1. A) Left: the edges between the 235 EEG electrodes. Right: the adjacency matrix $\boldsymbol{W}$. B) Left: the corresponding normalized Laplacian matrix $\breve{\mathcal{L}}$. Right: the eigenvalues of the normalized Laplacian matrix. C) Eigenvectors corresponding to eigenvalues λ_l with $l = 2$ (left) and $l = 221$ (right). Interpolation between vertices (i.e., EEG electrodes) was performed for visualization.

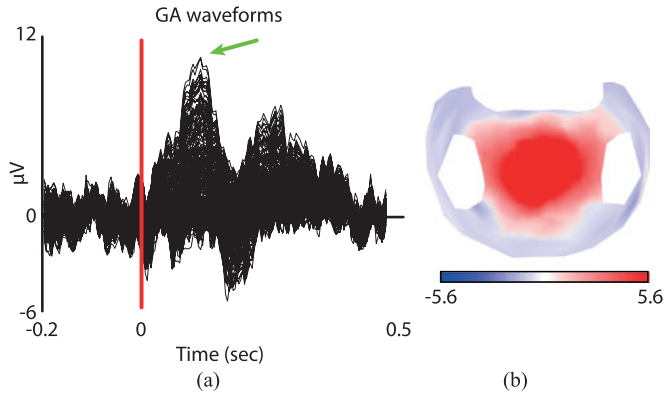

Fig. 2. A) GA waveforms of all electrodes time locked with the subject's response (red line) between 200-ms pre- and 500-ms postresponse. B) Topography (P3f topography) of the positive peak occurring 125-ms postresponse (i.e., t_{P3f} green arrow). Interpolation was performed for visualization.

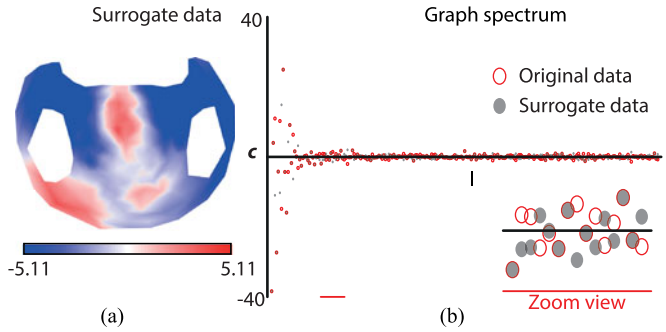

Fig. 3. A) An example of a surrogate data topography for the P3f topography. Interpolation was performed for visualization. B) Graph spectrum. Red and grey points indicate original and surrogate graph spectrum c, respectively. The inset shows a detailed view.

graph is real and symmetric, and, thus it has a complete set of orthonormal eigenvectors, $\boldsymbol{V} = [v_l]_{l=0,1,\ldots,N-1}$, that relate to nonnegative eigenvalues satisfying $\mathcal{L}\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda}$. The basis $\boldsymbol{V}$ formed by the eigenvectors of the graph Laplacian constitutes the GFT [1]. The eigenvalues can be interpreted as frequencies, and the eigenvectors as associated "oscillatory" signals. For spatial graphs that embed a regular grid with periodic boundary conditions, the conventional discrete FT is retrieved [5], [6]. For instance, the cycle graph leads to the equivalent of the one-dimensional FT; for each frequency, the complex exponential is split in two 90° phase-shifted eigenvectors (similar to a sin/cos pair) with the same eigenvalue (frequency) [10]. For more general graphs, these basis functions still correspond to our notion of FT; i.e., for larger eigenvalues, the oscillatory nature of the associated eigenvectors increases as well [6]. Since the eigenvectors are real-valued, the coefficients are also real-valued, and they only carry amplitude and sign.

Depending on the graph properties and on the application, it can be useful to consider alternative definitions of the graph Laplacian [5]. For instance, when the graph has a heterogeneous degree distribution, it is often convenient to consider the symmetrically normalized graph Laplacian: $\breve{\mathcal{L}} = \boldsymbol{D}^{-1/2}\mathcal{L}\boldsymbol{D}^{-1/2}$ where each weight $W_{i,j}$ is normalized by a factor of $1/\sqrt{D_i D_j}$ [1].

The GFT coefficients of a graph signal $\boldsymbol{x}$ are obtained by the projection $\boldsymbol{c} = \boldsymbol{V}^T\boldsymbol{x}$. The inverse transform corresponds to $\boldsymbol{x} = \boldsymbol{V}\boldsymbol{c}$. This unitary transform preserves energy (Parseval property).

III. Randomization in the Spectral Graph Domain

In the classical phase-randomization method, the amplitudes of the measured data's FT coefficients determine those of the surrogate data, which implies that the power spectral density of the surrogates is imposed as well as the autocorrelation in virtue of the Wiener–Khintchine theorem [3].

Similarly here, after transforming the measured graph signal into the spectral graph domain, we propose randomization by permuting or randomly generating the signs of the GFT coefficients $\boldsymbol{c}$. Next, the inverse GFT provides a realization of the surrogate graph signal. The graph equivalent of phase randomization preserves the amplitudes of the GFT coefficients and effectively imposes the surrogates' autocorrelation as defined via the graph Laplacian [9]. Specifically, under the null hypothesis, the measured graph signal is assumed part of the class of stationary signals with prescribed power spectral density or, equivalently, autocorrelation structure [9]. The null distribution of a test statistic can then be obtained from the surrogate graph signals and compared against its value for the measured signal.

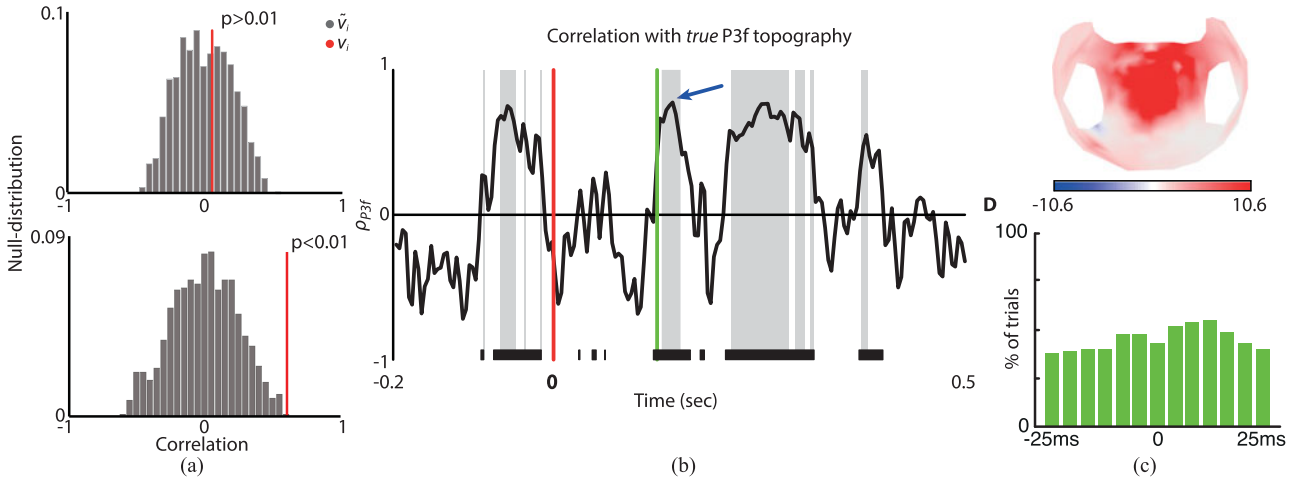

Fig. 4. A) Null distribution of the correlation coefficients for an EEG topography for which $\boldsymbol{\rho_{P3f}}$ is statistically not significant (Top) and significant (Bottom). The distribution is binned for visualization purposes. B) $\boldsymbol{\rho_{P3f}}$ for each time point of a single trial between 200-ms pre- and 500-ms postresponse. Red and green lines indicate the subject's response and the time $\mathbf{t}_{\mathrm{P3f}}$, respectively. Time points highlighted in grey are statistically significant using the proposed GFT sign-randomization, those in black using naïve spatial random permutation. C) Topography at the time point with the highest (significant) correlation (blue arrow in B). Interpolation was performed for visualization. D) Percentage of trials that have a statistically significant $\boldsymbol{\rho_{P3f}}$ for an interval around the $\mathbf{t}_{\mathrm{P3f}}$ (–25 ms to 25 ms). Zero corresponds to time $\mathbf{t}_{\mathbf{P3f}}$.

IV. Application to EEG

We chose high-density electroencephalography (EEG) to demonstrate the feasibility of the proposed method. We used a single-subject five-box visuospatial selective attention task dataset [11]. In the experiment, a monitor was showing five empty boxes arranged horizontally above the center of the screen that was placed in front of the subject. The subject was then instructed to press a response button whenever a disc appeared at the attended location.

We can build a spatial graph representation of the EEG cap (i.e., 235-electrodes cap) by associating each electrode with a vertex and encoding information of spatial neighborhood into the adjacency matrix [see Fig. 1(A)]. An edge between two vertices is present if the Euclidean distance between their three-dimensional coordinates does not exceed r:

$$W_{i,j} = \begin{cases} 1, & \text{if dist}\,(v_i, v_j) < r \\ 0, & \text{otherwise.} \end{cases}$$

We use $r = 4$ cm, which leads to a minimum/mean/maximum number of neighbors of 4/11/18. A similar scheme has been used to build a brain graph from magnetic resonance imaging (MRI) data [12]. We then use the eigendecomposition of the normalized Laplacian to define the GFT [see Fig. 1(B)]. Indeed, the eigenvectors associated with low eigenvalues are smooth and have for instance less zero-crossing than those with higher eigenvalues [see Fig. 1(C)].

Let us now consider a single EEG topography (i.e., reflecting the spatial distribution of the electrical potential) as a graph signal for which we can create surrogate signals with similar smoothness.

One common question in EEG analysis is to determine the similarity between a grand-average (GA) event-related-potential (ERP) topography (i.e., topography obtained by averaging single trial topographies) and topographies of single trials. This is important since the analysis of GA ERP features only fails to address the effects of behavioral tasks on the dynamics of cortical activity in single trials [11].

Here, we show that surrogate graph signals obtained using the proposed method can be employed to address this question with a stronger control of false positives than naïve spatial random permutation. The GA ERP topography is obtained by averaging the EEG data across 100 trials (time locked to the subject's response: –200 ms pre- to 500-ms postresponse) and by considering the time point t_{P3f} at which the GA waveforms present a positive peak [i.e., $t_{\mathrm{P3f}} = 125$ ms postresponse, see Fig. 2(A)]. The positive peak (P3f) had a large activation maximal at forehead sites [see Fig. 2(B)] in agreement with previous studies [13].

Using the proposed procedure with sign-randomization, we obtained $N_s = 999$ surrogate graph signals for the P3f topography, which allows determining p-values at a resolution of 0.1%. In Fig. 3(A), we show one realization of such a surrogate signal. By construction, the amplitudes of GFT coefficients for the original ($\boldsymbol{c}$) and the surrogate data ($\tilde{\boldsymbol{c}}$) are identical, but their signs differ [see Fig. 3(B)].

We can now obtain the null distribution of the Pearson correlation coefficient between an EEG topography at a given time point and trial, and the surrogate P3f topographies. Correlation with the *true* P3f topography (ρ_{P3f}) can then be considered significant when it is unlikely to occur according to this null distribution [see Fig. 4(A)]; i.e., when ρ_{P3f} exceeds the 99th percentile p_{99}, we reject the null hypothesis at α-level 1%, which is the probability of rejecting the null when it is true [14].

In Fig. 4(B), we report an example of a single-trial's correlation curve between instantaneous EEG topographies and the *true* P3f topography. The time points highlighted in gray are those at which the correlation is statistically significant using the proposed method (α-level = 1%). In black, significant correlation is indicated according to naïve spatial random permutation to

generate surrogates. As expected, this null hypothesis is much weaker and less realistic (i.e., it does not impose correlation structure), and thus it leads to more false positives at the same α-level. Also notice that at the moment of maximum correlation, the EEG topography presents a large positive activation at the forehead electrodes similar to the P3f topography [see Fig. 4(C)].

With our method, we can count the number of trials in which this similarity is significant in a given interval (e.g., 50 ms around t_{P3f}) by repeating the procedure with all the trials [$N_t = 100$, see Fig. 4(D)]. The majority of the trials (55%) have a positive frontal activation similar to that of the P3f topography 10 ms after t_{P3f}. Despite the decreased similarity over time, the frontal activation is retained in around 40% of the trials within the 50-ms interval around t_{P3f}. These results demonstrate that the frontal positive peak is preserved across single repetitions [13] showing the value of surrogate graph signals in assessing single-trial responses.

Furthermore, extensions of nonparametric testing in EEG analysis could include clustering of adjacent temporal samples such as proposed in [15].

V. Conclusion and Future Extensions

We proposed a novel framework for the generation of surrogate graph signals based on the GFT. Inspired by classical phase randomization of FT coefficients, we applied sign-randomization to the real-valued GFT coefficients. By construction, the power spectral density of the surrogate graph signals will be matched to spectral information of the original graph signal and as such maintain autocorrelation [9], [16].

We demonstrated a proof-of-concept with a high-density EEG dataset of 235 channels. Surrogate graph signals were generated that preserved "smoothness" of the reference EEG topography. We showed how the surrogate data could then be used to generate a null-distribution of correlation coefficients to evaluate the similarity between EEG topographical maps. We showed that this approach could contribute to analyze the effects of behavioral tasks on single-trial EEG topographies.

To conclude, we highlight a few extensions that could further increase the applicability of the proposed method.

A. Amplitude Adjustment

The amplitude histogram of the surrogate graph signal is not guaranteed to match the one of the original signal. Therefore, as in the original method [3], it is possible to adjust the amplitudes after reconstruction of the sign-randomized GFT coefficients using histogram matching.

B. Increased Randomization

Sign-randomization of real values has less degrees-of-freedom than phase randomization of complex values. However, it is possible to increase the amount of randomization by considering coefficients with close eigenvalues; i.e., GFT coefficients corresponding to such a group of *K* eigenvalues can be randomized using an arbitrary *K*-dimensional unitary transform.

C. Time-Dependent Graph Signals

In the case of time-dependent graph signals, the method can be combined with conventional phase randomization in the temporal Fourier domain. Similar schemes for spatiotemporal gridded data have been proposed for functional MRI [17]. In this case, the surrogate $\tilde{\boldsymbol{x}}$ would be obtained from the original time-dependent graph signal $\boldsymbol{x}$ ($N \times T$) as

$$\tilde{\boldsymbol{x}} = \boldsymbol{V}\boldsymbol{S}_S\boldsymbol{V}^T\boldsymbol{x}\boldsymbol{F}\boldsymbol{S}_T\boldsymbol{F}^T$$

where $\boldsymbol{S}_S \in \mathbb{R}^{N\times N}$ is a diagonal matrix encoding sign-randomization of the GFT coefficients; $\boldsymbol{S}_T \in \mathbb{R}^{T\times T}$ is a diagonal matrix encoding phase randomization of the temporal Fourier coefficients [3], [4]; $\boldsymbol{F}$ and $\boldsymbol{F}^T$ indicate temporal FT and IFT, respectively. Note that the sign-randomization of the GFT coefficients is the same over time in order to preserve the spatiotemporal correlation structure of the measured signal.

References

[1] D. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, "The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains," *IEEE Signal Process. Magazine*, vol. 30, no. 3, pp. 83–98, May 2013.

[2] S. Chen, R. Varma, A. Sandryhaila, and J. Kovacevic, "Discrete signal processing on graphs: Sampling theory," *IEEE Trans. Signal Processing*, vol. 63, no. 24, pp. 6510–6523, Aug. 2015.

[3] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. D. Farmer, "Testing for nonlinearity in time series: The method of surrogate data," *Physica D: Nonlinear Phenomena*, vol. 58, pp. 77–94, 1992.

[4] D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Phys. Rev. Lett.*, vol. 73, pp. 951–954, 1994.

[5] J. Irion and N. Saito, "Applied and computational harmonic analysis on graphs and networks," in *Proc. SPIE Opt. Eng. Appl.*, 2015, pp. 95971F-1–95971F-15.

[6] D. I. Shuman, B. Ricaud, and P. Vandergheynst, "Vertex-frequency analysis on graphs," *Appl. Comput. Harmonic Anal.*, vol. 40, pp. 260–291, 2016.

[7] D. K. Hammond, P. Vandergheynst, and R. Gribonval, "Wavelets on graphs via spectral graph theory," *Appl. Comput. Harmonic Anal.*, vol. 30, pp. 129–150, 2011.

[8] F. G. Meyer and X. Shen, "Perturbation of the eigenvectors of the graph laplacian: Application to image denoising," *Appl. Comput. Harmonic Anal.*, vol. 36, pp. 326–334, 2014.

[9] N. Perraudin and P. Vandergheynst, Stationary signal processing on graphs. *arXiv preprint, arXiv:1601.02522*, 2016. [Online]: Available: arxiv.org

[10] N. Leonardi and D. Van De Ville, "Tight wavelet frames on multislice graphs," *IEEE Trans. Signal Process.*, vol. 61, no. 13, pp. 3357–3367, Apr. 2013.

[11] A. Delorme, M. Westerfield, and S. Makeig, "Medial prefrontal theta bursts precede rapid motor responses during visual selective attention," *J. Neurosci.*, vol. 27, pp. 11949–11959, 2007.

[12] H. Behjat, N. Leonardi, L. Sörnmo, and D. Van De Ville, "Anatomically-adapted graph wavelets for improved group-level fMRI activation mapping," *NeuroImage*, vol. 123, pp. 185–199, 2015.

[13] S. Makeig *et al.*, "Electroencephalographic brain dynamics following manually responded visual targets,"" *PLoS Biol.*, vol. 2, pp. 747–762, 2004.

[14] B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," *Statis. Sci.*, vol. 1, pp. 54–75, 1986.

[15] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG-and MEG-data," *J. Neurosci. Methods*, vol. 164, pp. 177–190, 2007.

[16] C. Chatfield, *The Analysis Of Time Series: An Introduction*. Boca Raton, FL, USA: CRC press, 2013.

[17] R. S. Patel, D. Van De Ville, and F. D. Bowman, "Determining significant connectivity by 4D spatiotemporal wavelet packet resampling of functional neuroimaging data," *NeuroImage*, vol. 31, pp. 1142–1155, 2006.