

**Methods**

We adapt p-value combiners to work for frequency domain data.

A two-step procedure for graphical modelling on EEG data is proposed.

We control for false detections across networks with various level adjustments.

Methods are proposed to combine results across subjects and create group results.

Coloured graphs are made for each group, showing the prevalence of connections.

---

**ARTICLE INFO**

**Article history:**
Received 22 February 2016
Received in revised form 24 June 2016
Accepted 18 July 2016
Available online 21 July 2016

**Keywords:**
Partial coherence
Multivariate time series
Multiple hypothesis testing
p-Value combiners
Electroencephalogram (EEG) data

---

**ABSTRACT**

**Background:** In the graphical modelling of brain data, we are interested in estimating connectivity between various regions of interest, and evaluating statistical significance in order to derive a network model. This process involves aggregating results across frequency ranges and several patients, in order to obtain an overall result that can serve to construct a graph.

**New method:** In this paper, we propose a method based on p-value combiners, which have never been used in applications to EEG data analysis. This new method is split into two aspects: frequency-wide tests and group-wide tests. The first step can be effectively adjusted to control for false detection rate.

**Results:** This two-step protocol is applied to EEG data collected from distinct groups of mental health patients, in order to draw graphical models for each group and highlight structural connectivity differences. Using the method proposed, we show that it is possible to reliably achieve this while effectively controlling for false connections detection.

**Comparison with existing method(s):** Conventionally, the Holm’s Stepdown procedure is used for this type of problem, as it is robust to type I errors. However, it is known to be conservative and prone to false negatives. Furthermore, unlike the proposed methods, it does not directly output a decision rule on whether to accept or reject a statement.

**Conclusions:** The proposed methodology offers significant improvements over the stepdown procedure in terms of error rate and false negative rate across the network models, as well as in terms of applicability.

---

**1. Introduction**

The understanding of connectivity in large-dimensional time series has been a topic of central importance in neurology, and more precisely in neurological imaging. The interest in these techniques is widespread across imaging techniques (EEG, Medkour et al., 2010; fMRI, Marrelec et al., 2006) and experimental works of various types (learning experiments, Fiecas and Ombao, 2014; motor skills, Mima et al., 2000; resting-state, Salvador et al., 2005).

One of the most important features of neurological data analysis is functional connectivity between parts of the brain: how do various regions of interest interact? For this purpose, graphical modelling of time series is an ideal tool.

A well-known contribution to this field is the frequency-domain approach exposed in Dahlhaus (2000). In this methodology, the data is transformed into the frequency domain, where its co-dependency structure is analysed via the partial coherences. The partial coherence measures the connection between two series after the removal of the linear effects of the remaining series. It is a function of frequency and can be used to reflect connectivity across any frequency range $\Omega \subseteq [0, f_0]$. It is derived from $S^{-1}(f)$, the inverse of the spectral matrix $S(f)$. Its use in network analysis and
related methodologies has become very widespread in the literature (Makhtar et al., 2014; Mima et al., 2000; Pohja et al., 2002). Once the partial coherences have been estimated, an important question arises concerning the statistical significance of the results. A zero-valued estimate would indicate that no connection is detectable between two variables at a given frequency. The standard approach is to test for the following at all frequencies \( f \) in range:

\[
H_0 : \text{partial coherence} = 0 \text{ vs } H_1 : \text{partial coherence} > 0.
\]

In some cases, multiple measurements of the data are available, hence allowing the estimation of a bootstrapped distribution (Fiecas and Ombao, 2011). When this is not possible, the use of analytical tools is required. A popular alternative involves the Partial Mutual Information (PMI), which integrates the estimates at all frequencies within a band into one variable for each partial coherence, and is then compared to a threshold (Salvador et al., 2005). However, its distribution and statistical properties remain largely unknown, and it is often unclear how the threshold value should be set for a given band in \([0, f_H]\).

In this paper, we consider an alternative protocol that follows the method detailed in Medkour et al. (2009). Under conventional pre-processing and estimation methods for the spectral matrix \( S(f) \), its distribution and the distribution of its partial coherence are known (Goodman, 1963), and then, under \( H_0 \), the partial coherence estimates can be modelled as Beta-distributed random variables.

In the frequency domain approach, results are derived at each frequency \( f \), for all frequencies in range \( f \in \mathcal{F} \). This is many steps away from an overall graphical model. Once the partial coherences have been measured and tested for significance at a suitable level \( \alpha \), how can the results be aggregated across frequencies and subjects to deliver one graph?

This can be regarded as a multiple hypothesis testing problem, otherwise known as a conjunction analysis problem in Neurology (Friston et al., 1999). A traditional approach to this is the Holm’s Stepdown procedure (Lehmann and Romano, 2005; Holm, 1979), where the null hypothesis \( H_0 \) is tested for at each frequency ordered by \( p \)-values \( p(1) \leq p(2) \leq \cdots \). Every time \( H_0 \) is rejected, the procedure moves on to the next frequency, and only stops at the frequency \( L \) where \( H_0 \) is finally accepted. This method is robust to Type I errors, as it is designed to control the family-wise error rate (FWER) below a desired level \( \alpha \) (Lehmann and Romano, 2005). However, the stepdown is prone to false negatives. Furthermore, translating the resulting \( L \) into a decision is ambiguous. \( L \) may sometimes be very small, especially compared to the number of frequencies it is computed for. For instance, if \( L = 1 \) for some pairs of variables, should the connections between these pairs still be included in the graphical model? No substantial research has been carried out to answer this beyond some case-specific solutions (Medkour et al., 2010).

Multiple hypothesis testing is not limited to the use of the Stepdown procedure. In other applications, the use of \( p \)-value combiners is very prevalent. In Genomics, the Westfall–Young min-p procedure (Westfall and Young, 1993) proves to be very popular, as it is robust to Type I errors and can also handle correlated data by estimating the joint \( H_0 \) distribution through resampling. Other well-known combiners rely on Bayesian inference, such as Efron’s empirical Bayes method (Efron, 2003), where prior probabilities are assigned on the proportion of null and non-null statements and the false discovery rate is evaluated empirically.

In the context of spectral domain analysis, closed-form analytical methods tend to be preferred, due to the large computational cost associated with performing calculations at each frequency. In this category of methods, The Fisher (Fisher, 1932) and Simes (Simes, 1986) combiners constitute popular examples, that are widely used in applications of Computational Statistics, such as Genetic Epidemiology (Sungho et al., 2009) and Biostatistics (Chen et al., 2014). They deliver a single scalar that can then be tested on well-defined distributions, in order to ascertain the significance of an overall proposition. The use of these \( p \)-value combiners has been relatively rare in graphical modelling of neurolgical data thus far. Conventionally, they require that the set of multiple tests is independent, which is almost never the case with frequency domain data. However it is possible to generalise their use for this specific application.

In this paper, we review various \( p \)-value combiners and assess their suitability for graphical modelling of EEG data compared to the Stepdown procedure. After reviewing some background results in frequency domain analysis and multiple hypothesis testing, we propose a two-step procedure to carry out graphical modelling on EEG data.

• We demonstrate how the classical \( p \)-value combiners can be used on a subset of the frequency range that only includes uncorrelated data, and evidence their performance on simulated data.
• Test combiners can also be used to ascertain the significance of a graphical model for a sample population. Using EEG measurements from three distinct groups of mental health patients, we demonstrate how we can aggregate the results of each patient in all groups in order to obtain a group-wide coloured graphs, which show the intensity of connections in each group.
• When combining results across a frequency range, each connection between a pair of channel is evaluated independently. In doing so, it is important to control for the detection of false positives when constructing individual graphs. We show how this can be managed using a false edge detection adjustment, for both low-dimensional and larger dimensional data.

2. Background – graphical modelling

Let \( \{ X_t \} \) be a 2nd order stationary vector time series, \( \{ X_t \} \in \mathbb{R}^p \), \( t \in \{0, \ldots, T-1\} \), with an associated spectral matrix \( S(f) \in \mathbb{C}^{p \times p} \). Many estimation procedure exist for \( S(f) \), here we choose the multitaper spectral estimate for its good statistical and analytical properties (Percival and Walden, 1993). It starts with a set of \( K \) orthogonal tapers, satisfying the following property:

\[
\sum_t h_{k,k} = 1, \quad k \leq K.
\]

There are many types orthogonal tapers \( \{ h_{k,k} \} \) that are regularly used in the literature on spectral estimation, we choose here the Sine tapers, for their ease of implementation (Walden et al., 1995), defined for all \( k \leq K \) and \( t \leq T - 1 \):

\[
h_{k,k} = \sqrt{2 \pi \int_1^T \sin \left( \frac{(k+1)\pi}{T} t \right) dt}, \quad t \leq T, \quad k \leq K.
\]

Using the multiple tapers \( \{ h_{k,k} \} \), we can define the following Fourier transforms \( J_k(f) \) for \( k < K \) on the data \( X_t \):

\[
J_k(f) = \Delta_t \sum_i h_{k,k} X_{k,k} e^{-i2\pi f_k \Delta_t} \in \mathbb{C}^{p \times 1},
\]

which in turn can be used to create an estimate for the spectral matrix \( S(f) \), called the multitaper estimate:

\[
\hat{S}(f) = \hat{S}^{mt}_K(f) = \frac{1}{R} \sum_{k=1}^K J_k(f) J_k^*(f).
\]
The matrix $S(f)$ is instrumental in evaluating the dependency structure between each pair of $p$ channels. Doing so, we aim to build a graph for the data set $\{X_i\}$:

$$G = (V, E), \quad V = \{i, i \leq p\}.$$ 

$E$ represents the set of edges that connects channels to each other. Because we are interested in forming an undirected, conditional correlation graph, we define the set of edges $E$ as follow:

$$(i, j) \not\in E \Rightarrow X_i \perp X_j | X_{\{i,j\}}, \quad \forall i, j \leq p, \ i \neq j.$$ 

The term “$X_i \perp X_j | X_{\{i,j\}}$” defines conditional independence between channel $i$ and $j$ – in other words, no direct connections between these two channels when excluding the $p-2$ others. This can be reformulated using the partial coherence (Dahlhaus, 2000; Medkour et al., 2009):

$$\gamma_{\{i,j\}}^2(f) = \frac{(S^H(f))^2}{S^H(f)S^H(f)}, \quad S^H(f) = (S^{-1}(f))_{ij},$$

$$(i, j) \not\in E \Rightarrow \gamma_{\{i,j\}}^2(f) = 0 \quad \forall f \in [0, f_N].$$

Hence, we bring our attention to the partial coherences for the construction of the graph $G$. Because it relies on the unknown quantity $S^{-1}(f)$, we use estimates for the partial coherences:

$$\hat{\gamma}_{\{i,j\}}^2(f) = \frac{(\hat{S}^H(f))^2}{\hat{S}^H(f)\hat{S}^H(f)}, \quad \hat{S}^H(f) = \hat{S}^{-1}(f),$$

$$\hat{S}^{-1}(f)$$ is a suitable estimator for $S^{-1}(f)$, derived from $S^{(mt)}(f)$. In order to construct the set of edges $E$ in the graph $G$, we need to test for the significance of the partial coherence estimates at every pair $(i, j)$:

$$H_0 : \gamma_{\{i,j\}}^2(f) = 0 \quad \forall f \in \Omega \quad vs \quad H_1 : \gamma_{\{i,j\}}^2(f) > 0 \quad \forall f \in \Omega \quad (2.3)$$

For each pair of channels $i$ and $j$, we derive a decision on whether to accept/reject $H_0$ at level $\alpha$, using the protocol in Section 3. This can be translated into a graphical model as follow:

$$\hat{G} = (\hat{V}, \hat{E}).$$

Accept $H_0 \Rightarrow (i, j) \notin \hat{E}$,

Reject $H_0 \Rightarrow (i, j) \in \hat{E}$.

We then have an estimate $\hat{G}$ of the true graph $G$ for the data $\{X_i\}$.

This testing procedure is repeated for all $n_{pairs} = \binom{p}{2} = p(p-1)/2$ pairs $(i, j), \ i < j \leq p$, until the set of edges $\hat{E}$ is fully established. Each edge is tested independently, as they are defined as the conditional dependence of different pairs of channels while discounting the others. The risk of detecting a false edge increases with $n_{pairs}$. Therefore, the probability of including an edge in $E$ erroneously should be controlled, as covered in Section 5. We propose two methods to achieve this, on that is strict and the other more permissive, which will be appropriate depending on the dimensions of the data.

The same applies for population-wide testing. Consider a set of $k \leq N$ data samples $\{X_i^k\}$, collected from $N_G$ distinct groups of subjects, with $N_G$ individuals in each group. We may be interested in estimating a group-wide graphical model $G^k = (V, E^k)$ for each group $g \leq N_G$, starting with individual graphical models for each $i \leq N_G$:

$$\forall i \leq N_G, \quad \hat{G}^i = (\hat{V}, \hat{E}^i).$$

The individual graphs $\hat{G}^i$ need to be combined for each group in order to derive a group-wide estimate $\hat{G}^k$ for the true graph $G^g$ of the population $g$.

This echoes the questions covered by multi-subject conjunction analysis (Friston et al., 1999; Heller et al., 2007), where in a group of subjects we need to ascertain the proportion $u$ of individuals who exhibit a trait – in our case a connection $(i, j) \exists i \neq j \leq p$. While in general multiple hypothesis testing, we would test for $u > 0$, in conjunction analysis we want to test for a range of $u \in [0, 1]$:

$$H^g_0 : k < u \quad vs \quad H^g_1 : k \geq u,$$

where $k$ is the true proportion of individuals in the group $g$ exhibiting the trait of interest. The hypothesis test of Eq. (2.5) can be repeated for a range of $u$’s, and the results of these combined in a “coloured” graph, where a colour is associated with each value of $u$. This is covered in Section 6.

3. Hypothesis tests for graphical modelling

Eq. (2.3) can be evaluated using multiple tests $H_0(f)/H_1(f)$ at each frequency $f$, and subsequently aggregated:

$$H_0(f) : \gamma_{\{i,j\}}^2(f) = 0 \quad vs \quad H_1(f) : \gamma_{\{i,j\}}^2(f) > 0, \quad (3.1)$$

discretised frequency $f \leq f_N$. Under $H_0(f)$, when using the multitaper estimate of Eq. (2.1) with $K > p$, the estimate $\hat{\gamma}_{\{i,j\}}^2(f)$ of Eq. (2.2) is known to have a Beta$(1, K-p+1)$ distribution (Medkour et al., 2009):

$$\hat{\gamma}_{\{i,j\}}^2(f) \sim B(1, K-p+1) \quad (3.2)$$

An important consideration is that the data at each frequency is not always independent of other frequencies, depending on the bandwidth of the spectral estimate.

**Definition 3.1.** The bandwidth of a spectral matrix estimate is the spectral window covered by said estimator, and is analogous to its resolution. In our case, using a multitaper estimate with sine tapers, we have (Walden et al., 1995):

$$B := \frac{K + 1}{\Delta f(N+1)} \quad (3.3)$$

The bandwidth of the data in applications is defined by $\Delta f = f_{i+1} - f_i$, for any $1 \leq i \leq N_f, N_f$ representing the number of discretised frequencies in $[0, f_N]$. If $B < \Delta f$, the data across frequencies is deemed independent, otherwise we consider that there exists dependency between frequencies.

**Remark 1.** Data at any frequencies $f_i, f_j \in [0, f_N]$ such that $|f_i - f_j| > B$ are independent (Percival and Walden, 1993).

The nature of the correlation structure of the spectral data across various frequencies is not well established, but it can be assumed that it is strictly non-negative:

$$\forall f_i, f_j, \mbox{s.a.} \ |f_i - f_j| \leq B, \quad \mbox{Cov}(\hat{S}(f_i), \hat{S}(f_j)) \geq 0$$

**Remark 2.** Spectral estimates reliant on smoothing/tapering are virtually never free of inter-frequency correlation, as the smoothing process incurred by the smoothing/tapering widens the bandwidth $B$. In spite of this, this class of spectral estimators is still preferred for their other statistical and inversion properties.

With every hypothesis test procedure, we pay close attention to the levels of the tests $\alpha$. It represent the upper limit we are willing to tolerate on the probability of making a false detection, or in other words, the probability of making a type I error:

$$\forall (H_0(f) \mbox{ rejected} | H_0(f) \mbox{ true}) < \alpha.$$
Across a set – or a “family” – of $\mathcal{T}$ multiple and/or simultaneous tests, it is equally important to consider the family-wise error rate (FWER);

**Definition 3.2 (Family-wise error rate (FWER), Lehmann and Romano, 2005).** The family-wise error rate, in the context of $\mathcal{T}$ multiple tests $H_1^0, H_1^1, \ldots, H_T^0, H_T^T$, is the probability of making one or more Type I errors across all the tests:

\[
\text{FWER} := \mathbb{P}(|V| \geq 1),
\]

\[
V = \{i \leq \mathcal{T} | H_i^0 \text{ accepted, } H_i^0 \text{ true}\}
\]

where $|V|$ represents the size of the set $V$.

Hence, FWER control is an extension of Type I error control to the context of multiple hypothesis testing. Thus, for a given level $\alpha$, we require that:

\[
\text{FWER} \leq \alpha.
\]

An important related concept is the false discovery rate (FDR).

**Definition 3.3 (False discovery rate (FDR), Lehmann and Romano, 2005).** The FDR is the expected rate of Type I errors in a multiple hypothesis testing:

\[
V = \{i \leq \mathcal{T} | H_i^0 \text{ accepted, } H_i^0 \text{ true}\},
\]

\[
\mathcal{T}_0 = \{i \leq \mathcal{T} | H_i^0 \text{ true}\},
\]

\[
\text{FDR} = \mathbb{E}\left(\frac{|V|}{\mathcal{T}_0}\right)
\]

**Remark 3.** It can easily be shown that:

\[
\text{FDR} \leq \text{FWER},
\]

with equality holding if $H_0$ is true, and otherwise the inequality is strict. Thus, any form of FDR control is more liberal than FWER control, in the sense that it will permit more rejections (Lehmann and Romano, 2005).

Equally, the Bonferroni inequality is a central result in the understanding of FWER/FDR control.

**Definition 3.4 (Bonferroni Inequality, Lehmann and Romano, 2005).** For a set of $\mathcal{T}$ hypothesis tests with $p$-values $p_1, \ldots, p_\mathcal{T}$, we have:

\[
\mathbb{P}(\cup_{i=1}^\mathcal{T} (p_i \leq \alpha/\mathcal{T})) \leq \alpha.
\]

The implication of this result is that the FWER can be controlled while testing for $H_0 : H_0^0 \text{ true } \forall i \leq \mathcal{T}$ vs $H_1 : H_i^0 \text{ true } \exists i \leq \mathcal{T}$ at significance level $\alpha$, if we test for each of the $\mathcal{T}$ individual test at level $\alpha/\mathcal{T}$. While this result is very rich, it can be too strict when $\mathcal{T}$ becomes large, and induce large type II error rates. Throughout the rest of the paper, we rely on this result:

**Lemma 3.5 (Dickhaus, 2014, Thm 2.1).** For multiple hypothesis tests $H_1^0 \text{ vs } H_1^1$ with $p$-value $p_i$ where the test statistics have a continuous distribution, the $p$-values are uniformly distributed under the general null hypothesis $H_0$:

\[
p_i | H_0 \sim U(0, 1)
\]

Having established the nature of the set of hypothesis tests we want to run, we now review the various protocols available to adequately perform the multiple hypothesis tests in Section 4.

4. **$p$-Value combiners**

Multiple hypothesis tests can be handled with various $p$-value combiners. Here we present three approaches: the Holm’s Stepdown procedure, the Sime’s approach and the Fisher combiner. The first can handle correlated tests and returns a scalar $L$ representing the number of tests where the statement was significant, and is also a methodology that is widely used in multiple hypothesis testing applied to neurological data. The other two return a single $p$-value, but only work for independent tests.

4.1. **Holm’s Stepdown procedure (Holm, 1979)**

Multiple hypothesis testing, especially in the case of correlated test, can be approached with the Stepdown procedure. In the context of Eq. (2.3), we have a set of $\mathcal{T} = N_f$ hypothesis tests to perform on the data, with decisions $H_0^0, H_1^1$ for each $i \leq N_f$, an associated test statistics $T_i$ and $p$-value $p_i$. Denote the ordered test statistics and $p$-values as follow:

\[
T_i^{(1)} \geq \cdots \geq T_i^{(N_f)}, \quad p_i^{(1)} \leq \cdots \leq p_i^{(N_f)},
\]

and their corresponding hypotheses as $H_0^{(1)}, \ldots, H_0^{(N_f)}$.

In the Stepdown approach, we test sequentially for a subset of the multiple hypothesis:

\[
\mathcal{H}_0^i : \{H_0^{(1)}, \ldots, H_0^{(N_f-i+1)}\}, \quad \mathcal{H}_0^i \text{ accepted}
\]

\[
\text{for } j = \arg\min_i \{T_i^{(N_f-i+1)} \leq C_i\}.
\]

The Holm’s Stepdown procedure can be used on the hypothesis test of Eq. (2.3) and on the conjunction analysis tests of Eq. (2.5) in the following way:

\[
C_i = \frac{\alpha}{N_f - i + 1}, \quad \text{Accept } \mathcal{H}_0^i \text{ for } j = \arg\min_i \{p_i^{(j)} \geq C_i\}. \quad \text{(4.1)}
\]

In either form, the Stepdown procedure delivers a scalar $L$:

\[
L = \arg\min_j \{\mathcal{H}_0^j \text{ accepted}\}.
\]

When applied to frequency-wide data, this shows the number of discretised frequencies $f \leq N_f$ for which the estimated partial coherences were significantly greater than zero. This protocol is uniformly more powerful than just applying a Bonferroni correction based on Definition 3.4, as it is more capable of detecting cases where $H_0$ is false (Holm, 1979; Lehmann and Romano, 2005). For the purpose of frequency-wide testing, we use thresholding and reject $H_0$ in Eq. (2.3) if $L > 0$.

When applied to subject conjunction, this shows the number of subjects in a group for whom a trait was deemed significant.

For the following methods, we only work across uncorrelated frequencies, which are $B$ Hz apart (Eq. (3.3)). Hence, we have $M = \lfloor N_f / B \rfloor$ eligible data entries and $\mathcal{T} = M$ multiple tests.

4.2. **Sime’s modified Bonferroni approach**

Sime’s method builds directly on the Bonferroni Inequality, and works for independent tests. Although it has been shown to apply to positively correlated tests as well, we will not recouse to this extension here, as it isn’t clear whether partial coherences estimates between frequencies are non-negatively correlated.

**Lemma 4.1 (Simes, 1986).** Let $p_1 \leq p_2 \leq \cdots \leq p_M$ be the ordered $p$-values associated with $M$ multiple tests, and all $p_i$ follow the $U(0, 1)$ distribution under $H_0$. Set a test procedure that decides to reject $H_0$ when there are some $j \leq M$ such that:

\[
p_{(j)} \leq \alpha \times j / M.
\]

This test procedure has a FWER $= \alpha$ when the multiple tests are independent.
**Lemma 4.1** is a consequence of the Bonferroni inequality, as shown in Simes (1986).

From Lemma 4.1 we can reformulate the general null hypothesis \( H_0 \) as follow:

\[
(p_{ij}) \equiv \{p_{ij} \mid j > \alpha, \forall j \leq N_I\} \quad \text{and} \quad \{p_{ij} \mid j > \alpha, \forall j \leq N_I\} = \left\{ M \times \min_{j} \left( \frac{p_{ij}}{M} \right) > \alpha \right\}.
\]

Based on this, we can introduce a \( p \)-value combiner built on **Lemma 4.1**:

\[
p_{\text{Simes}} := M \times \min_{j \in \mathbb{M}} \left( \frac{p_{ij}}{M} \right)
\]

(4.2)

\[
P_{\text{Simes}}|H_0 \sim U(0, 1)
\]

\( H_0 \) is then rejected in Eq. (2.3) if \( P_{\text{Simes}} < \alpha \) in Eq. (4.2).

### 4.3. Fisher \( p \)-values combiner

The Fisher combiner (Fisher, 1932) is by far the most popular choice of \( p \)-value combiners for independent tests, as it has a well-defined formulation and distribution under \( H_0 \):

\[
F = -2 \sum_{i=1}^{M} \log(p_i),
\]

\[
F|H_0 \sim \chi^2_{2M}.
\]

We reject \( H_0 \) in Eq. (2.3) if \( F > C(\chi^2_{2M}, \alpha) \), where \( C \) represents the critical value of the \( \chi^2_{2M} \) distribution at level \( \alpha \).

### 5. Network-wide FDR control

In the previous sections, we have covered various techniques to combine multiple tests across a set of discretised frequencies in order to deliver a decision rule for the whole frequency range \( \Omega \). This is done for each pair of channels \( (i, j), i < j \leq p \). Doing so, we also need to control for false positives across pairs of connections when evaluating individual graphs.

The above procedure is repeated for all \( n_{\text{pairs}} \) pairs \( (i, j), i < j \leq p \) with \( n_{\text{pairs}} = p(p - 1)/2 \) — until the set of edges \( \hat{E} \) in Eq. (2.4) is fully established. Each edge is tested independently, as they are defined as the conditional dependence of different pairs of channels while discounting the others. This in itself defines a new family of tests, and as such the risk of detecting a false edge increases with \( n_{\text{pairs}} \):

\[
\pi(\exists i,j, (i, j) \in \hat{E}(i, j) \notin E) \leq \sum_{k} \pi(k = (i, j) \in \hat{E}(k \notin E) = n_{\text{pairs}} \pi(k = (i, j) \in \hat{E}(k \notin E) = n_{\text{pairs}} \alpha
\]

(5.1)

This directly relates to the Bonferroni inequality (Definition 3.4). Hence, an appropriate course of action when testing for the \( n_{\text{pairs}} \) edges in the graph \( \hat{C} \) of Eq. (2.4) is to set the significance level at \( \alpha' = \alpha/n_{\text{pairs}} \) instead of just \( \alpha \) in order to minimise the probability of detecting false edges. For each pair of variables \( (i, j) \), denote their associated \( p \)-values \( p_{(i,j)} \) output from the combination of results across frequencies. We then have the following rule:

\[
(i, j) \notin \hat{E} \text{ for all pairs } j = \alpha^*.
\]

(5.2)

This correction becomes very strong as the dimensions \( p \) grow large. For instance, with \( p = 50 \) dimensions, this would lead to a corrected level \( \alpha^* = 4 \times 10^{-3} \) from \( \alpha = 0.05 \), rendering the detection of truly-existent edges almost impossible. An alternative is to allow some relaxation in the correction required on the level \( \alpha^* \), without losing out too excessively on the FDR control.

In the context of conjunction analysis, the Sime’s combiner can be used in the hypothesis test of Eq. (2.5). Building on a protocol detailed in Benjamini and Yekutieli (2001) and in (Heller et al., 2007), we propose the following methodology:

\[
k = \max_{j} \left( \frac{p_{ij}}{\min_{j} \alpha} \right).
\]

(5.3)

\( H_0 \) is rejected if \( k > u_{\alpha} \). This is in fact also known as the Benjamini and Hochberg step-up procedure.

Denote the ordered \( p \)-values of each connection as \( p_{(1)} \leq \cdots \leq p_{(n_{\text{pairs}})} \), we make use of Eq. (5.3) and create the following rule:

\[
\alpha^*(k) = \alpha \times k/n_{\text{pairs}},
\]

(5.4)

\[
(i, j) \notin \hat{E} \text{ for all pairs } j = \alpha^*(k).
\]

**Lemma 5.1.** Under non-negative dependency, the procedure in Eq. (5.4) controls the FDR.

**Proof.** This topic and associated proofs are extensively covered by Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). □

**Remark 4.** The Benjamini and Hochberg step-up procedure is comparable to the Holm’s Stepdown procedure of Section 4.1, but aims to control the FDR while the Stepdown procedure performs a strict FWER control.

Therefore, the procedure of Eq. (5.2) is a form of FWER control, whereas the procedure of Eq. (5.4) is a type of FDR control. Based on Remark 3, we expect the second scheme to be less conservative than the first.

### 6. Multi-subject conjunction/group-wide graphical modelling

Thus far, the various testing protocols on the discretized frequencies and connections of channels have been performed for each individual in the study, and some attention has been paid on FDR control for the purpose of evaluating individual graph models. When dealing with groups of subjects, the next stage is to aggregate these results across all subjects in order to obtain a unique group-wide graph.

Consider a collection of \( k \leq N \) time series data \( \{X_{g}^{i}\} \), split into \( N_{C} \) distinct groups. We are interested in estimating a graphical model \( G = (V, E) \) for each group \( g \leq N_{C} \). In each of these groups, there are \( N_{g} \) individuals — such as \( \sum_{g} N_{g} = N_{C} \) — for which we estimated a graphical model from the \( p \)-values associated with each edge \( p_{(1)}, \ldots, p_{(n_{\text{pairs}})} \):

\[
\forall i \leq N_{g}, \tilde{G}_{g} = (V, \tilde{E}), \quad \tilde{E} = f(p_{(1)}, \ldots, p_{(n_{\text{pairs}})}).
\]

for some \( p \)-value combiner function \( f(\cdot) \). The \( p \)-values \( p_{(1)}, \ldots, p_{(n_{\text{pairs}})} \) need to be aggregated across all individuals \( i \leq N_{g} \) in each group in order to derive a group-wide graphical model estimate \( \tilde{G}_{g} \). We do so via some function \( f(\cdot) \) that takes the \( p \)-values of all the individuals \( i \leq N_{g} \) in the group \( g \) as inputs.

We are interested in assessing whether \( \tilde{G}_{g} \) is the true graph \( G^{g} \) of the group \( g \). In doing so, we are concerned with controlling for false edge detection between any two channels \( j \) and \( k \):

\[
p_{j,k,g}^{g} = P\{ (j, k) \in \tilde{G}_{g} | (j, k) \notin G^{g} \}.
\]
We denote the proportion of individuals in group g for whom the connection (j, k) is deemed significant as follow:

\[ P_{jk} = \frac{|\{i \in [1, N_g]: (j, k) \text{ significant for } i\}|}{N_g}, \]  

(6.1)

Note that \( P_{jk} \) is not equal to the proportion of graph estimates \( \hat{C}^g \) that exhibit the connection (j, k):

\[ P_{jk} \neq \frac{|\{i \in [1, N_g]: (j, k) \subseteq \hat{C}^g\}|}{N_g}. \]

The quantity \( P_{jk} \) can be computed using:

- The Stepdown procedure (Eq. (4.1)), where the output L represents the number of individuals in group g for which the connection (j, k) is significant. The quantity \( P_{jk} \) would then be equal to \( P_{jk} = \frac{1}{N_g} \).
- The Benjamini & Hochberg procedure (Eq. (5.4)), where for each connection (j, k), in the group g, denote the p-value \( p \) associated with individual i, and the ordered p-values \( p^{(1)} \leq \cdots \leq p^{(N_g)} \), and set

\[ u^* = \arg \max_{i \in N_g} (p_i \leq \frac{\alpha}{N_g}), \]

then \( P_{jk} = \frac{u^*}{N_g} \).

**Remark 5.** The Stepdown and Benjamini & Hochberg procedures are effectively a form of group-wide FDR control, equivalent to the procedures proposed in Section 5 for the making of individual graphs.

We propose to construct the graph \( \hat{C}^g \) based on the individual p-values \( p_i \), \( i \in g \), for each edge, using the subjects proportions \( P_{jk} \). For any connection (j, k), given a threshold level \( \theta \in (0, 1) \),

\[ P_{jk} \geq \theta \rightarrow (j, k) \in \hat{C}^g, \]

\[ P_{jk} < \theta \rightarrow (j, k) \notin \hat{C}^g. \]

In this context, the estimate \( \hat{C}^g \) is in fact a function of \( \theta \), i.e. \( \hat{C}^g = \hat{C}^g(\theta) \). Here, \( \theta \) represents the minimum required proportion of subjects that the channels connections need to be significant for, in order to include them in a graphical model estimate \( \hat{C}^g(\theta) \). This procedure can be repeated for any value of \( \theta \in [0, 1] \). Thus, for each group g, estimates \( \hat{C}^g(\theta) \) can be devised for a range of \( \theta \)'s in \( \Theta \) and overlaid to form a general graphical model estimate:

\[ \theta \in \Theta \subseteq \{i/N_g|i=1, \ldots, N_g\}. \]

This creates a “coloured” graph.

**Definition 6.1 (Coloured Graphs).** A coloured graph is a graphical model \( \hat{G} \) constructed from several sub-graphs \( \hat{G}^g(\theta) \), derived from thresholding the variables \( P_{jk} \) for a set of several \( \theta \in \Theta \):

\[ \hat{G} = \bigcup_{\theta \in \Theta} \hat{C}^g(\theta). \]

Then, each \( \theta \in \Theta \) can be associated with a colour – or a “line-style” – and each edge (j, k) inside of \( \hat{G} \) is associated with the colour of the largest \( \theta \) such that (j, k) \( \in \hat{G}^g(\theta) \).

For each connection in the graph and a range of thresholds \( \theta \)'s, an edge is included in a graphical model if it is significant for any threshold \( \theta \in \Theta \):

if \( P_{jk} \geq \theta \), then \( (j, k) \in \hat{G}^g(\theta) \),

\[ \hat{G} = \bigcup_{\theta \in \Theta} \hat{C}^g(\theta). \]

(6.2)

**Remark 6.** For any two thresholds \( \theta, \theta' \) in [0, 1] such that \( \theta > \theta' \) and any pair of variables \( (j, k) \), if \( (j, k) \in \hat{G}^g(\theta) \) then it follows that \( (j, k) \in \hat{G}^g(\theta') \).

In the context of building a graphical model with various \( \theta \in \Theta \), a colour is associated with each \( \theta \) and matched to an edge in \( \hat{G}^g(\theta) \) if it belong to \( \hat{C}^g(\theta) \).

This colouring system shows how prevalent connections are in the graphical models for each group in terms of subjects proportion. This type of representation is highly appropriate for neurological data analysis, as it allows to create intensity maps to represent the strength of neural activities in the brain, in the style of Heller et al. (2007, p. 1184).

Alternatively, researchers can decide what choice is appropriate for a fixed value \( \theta = \hat{\theta} \), and estimate a graphical model using it on its own: \( \hat{G} = \hat{C}(\hat{\theta}) \). This choice will depend on the nature of the data being analysed and the level of inference required.

Using these methods, we can ensure that for a group g the probability \( P_{jk} \) is bounded above by \( \alpha \), for all connections (j, k).

7. Simulation results

We apply three p-value combiners to several VAR(1) data models, in order to assess their ability to retrieve the true network structure of each model:

- Holm’s Stepdown approach (Eq. (4.1)), applied as a dependent and independent scheme. In the dependent case, results from all frequencies are combined, whereas in the independent scheme only independent frequencies are considered (see details in Section 3).
- Simes combiner (Eq. (4.2)).
- Fisher combiner (Eq. (4.3)).

We look at two groups of models, that differ in their dimensions:

Group 1: Low to moderate dimensions, \( p = 10 \),

Group 2: Moderate to high dimensions, \( p = 35 \).

Both group consist of five VAR(1) models, all expressed in the following way:

\[ X_t = \Phi X_{t-1} + \eta_t, \]

with \( t \in [0, T-1] \), \( T = 3072 \), \( \eta_t \sim i.i.d. \mathcal{N}_p(0_{p 	imes 1}, I_p) \), \( \Delta t = 1 \), hence \( f_t = 0.5Hz \), generating \( N_p = 3073 \) discretised frequencies on the range [0, 0.5] Hz. Each VAR(1) model is stationary, and produced using the semi-random VAR(1) model generator detailed in Appendix A. The true spectral matrices \( \Phi(f) \) for each model is found as described in Appendix B. In each group, the various five models differ in their Degree of Connectivity (DoC), which represents how much the variables of a multivariate system are connected to one another.

**Definition 7.1 (Degree of Connectivity (DoC)).** For a multivariate time series \( X_t \in \mathbb{R}^{p \times p} \), and a graph \( G = (V, E) \) fitted to the data \( X_t \), we define the DoC to be the proportion of existing edges in the set of all possible connections:

\[ \text{DoC}(G) = \frac{|\{(i, j) \in E| \theta(i, j)\}|}{\#(E)}. \]

For instance, a set of perfectly independent p-dimensional variables \( X_t \in \mathbb{R}^{p \times p} \) would return a graph G with no edges, and a DoC of 0%, whereas a perfectly connected graph G would have a DoC of 100%.

The DoCs of the models under consideration are displayed in Table 1.

For each model in each group, we create \( m = 500 \) data-copies \( X''_t \). For each one of them, we compute the multitaper spectral estimate \( \hat{S}(f) \) (Eq. (2.1)) with \( K = 1.2 \times p \) – resulting in a bandwidth of \( B = 0.005 \) Hz for group 1, and \( B = 0.015 \) Hz for group 2. We subsequently apply a diagonal up-lift with a factor of \( \eta = 10^{-8} \) in order to
stabilise inversion. The estimated partial coherences are then calculated from the inverted spectral matrix estimate $S^{-1}(f)$ at each discretized frequency $f$, and tested according to the protocol in Eq. (3.2). We then aggregate the results for each partial coherence using each $p$-value combiner: Fisher, Simes, and Holms. The combiners are then transformed into critical values $C$, the $p$-value of each combiner under their $H_0$ distribution. This then produces an adjacency matrix:

$$A(C) = \{A_{ij}(C)\} = \begin{cases} 1 & C < \alpha \\ 0 & C \geq \alpha \end{cases}$$

In the case of the Holm’s Stepdown procedure, we set $C = 1$ and $A_{ij}(C) = 1$ if $L > 0$.

For consistency with the rest of the literature on the subject of hypothesis testing, we set $r_0 = 0.05$.

Using the adjacency matrix produced by each combiner, we can bring our attention to the average error rate (AER), which is measured in terms of Hamming distance between the estimated graph $\hat{G}$ with adjacency matrix $A(C)$ and the true graph $G$ with adjacency matrix $A$ (Banks and Carley, 1994; Medkour et al., 2010):

$$\text{Hamming}(\hat{G}, G) = \sum_{(i,j)} |A_{ij}(C) - A_{ij}|,$$

$$\text{AER} = \hat{p}_m \left( \text{Hamming}(\hat{G}, G) / p(p - 1)/2 \right),$$

where we have $\#(i,j) = \#(i,j) = p = p(p - 1)/2$ possible connections in the graph.

Alternatively, we also look at the Average False Positive (AFP) rate and the Average False Negative (AFN) rate, which are defined as the AER over the set of non-existent and existing connections respectively:

$$\text{AFP} = \hat{p}_m \left( \frac{1}{m_{EC}} \sum_{(i,j) \notin E} |A_{ij}(C) - A_{ij}| \right),$$

$$\text{AFN} = \hat{p}_m \left( \frac{1}{m_E} \sum_{(i,j) \in E} |A_{ij}(C) - A_{ij}| \right),$$

where we have:

$m_{EC} = \#(i,j) \notin E$

and $m_E = \#(i,j) \in E$.

**Remark 7.** The AFP rate in Eq. (7.3) is analogous to the FDR of Definition 3.3, whereas the AFN rate is a measure for type II errors.

Results are averaged over all $m$ copies and displayed in terms of the evolution of the AER (Eq. (7.2)), AFP rate and AFN rate (Eq. (7.3)) for each $p$-value combiner, as a function of DoC, in Fig. 1 for group 1, and in Fig. 2 for group 2. For all models, we evaluate edge detection using three approaches: with regular levels $\alpha = 0.05$, with Bonferroni-adjusted levels $\alpha = 0.05/n_{\text{pairs}}$ of Eq. (5.2), and with Benjamini & Hochberg adjusted levels $\alpha(k) = \alpha \times k/n_{\text{pairs}}$ of Eq. (5.4).

### 7.1. Main results

We can see for group 1 in Fig. 1 that, when using the level $\alpha = 0.05$ on the edges, the overall Average False Positive rate is not actually bounded above by 0.05 for any $p$-value combiner, even with the more conservative dependent Stepdown procedure. Indeed, the dependent Stepdown scheme consistently achieves lower AFP rates than its counterparts. The stepdown protocol being designed to perform a strict FWER control, this is expected. On the other hand, the Fisher combiner is known to exaggerate evidence against the null hypothesis in multiple tests, which is manifested here with higher AFP rates and lower AFN rates across most models.

In Fig. 1, the adjustments on the significance level $\alpha$ proposed in Section 5 manage to reduce the AFP rate of all combiners. The reduction in AFP rate is most significant with the Bonferroni adjustment of Eq. (5.2), $\alpha = 0.05/n_{\text{pairs}}$. This comes at the cost of higher AER and AFN rates for all combiners. The same observations can be made for the level adjustment of Eq. (5.4), $\alpha(k) = 0.05 \times k/n_{\text{pairs}}$, but the impact on the AER and AFN rates is much more moderate. This is expectable, as this type of level adjustment is known to be less strict than a Bonferroni correction. However the reduction in AFP rate it achieves is limited, and for low dimensions such as $p = 10$ it is only marginally better to using no correction at all.

The benefits of using the Benjamini & Hochberg (B&H) level correction of Eq. (5.4) are more obvious on larger dimensional data, for which the results are displayed in Fig. 2. We can see that the more conventional Bonferroni adjustment of Eq. (5.2) is far too strict, leading to all $p$-value combiners to return AFP rates ranging from 45% to 99%. The B&H level correction manages to reduce the AFP rate considerably for most combiners, even down to 20% for the Fisher combiner for the models with the higher DoCs. While this is a significant improvement over the conventional Bonferroni adjustment, this still represents a substantial level of false negatives in the estimated graphical models, especially compared to the results obtained on lower dimensional data in Fig. 1. This calls for further investigations of false positives control in graphical models of large dimensions.

In Figs. 1 and 2, the Fisher combiner stands apart from its counterpart in terms of all divergence criteria. While its AFP rate can be higher compared to other combiners, it manages an overall AER that is lower in most models, but especially so for highly connected VAR(1) models. More importantly, it is more robust to type II errors than its counterparts. This makes it an highly eligible $p$-value combiner for higher dimensional data, such as in Fig. 2, where none of the other methodology were able to return AFN rates below 60%. Based on its low AER and despite its higher AFP rate, the Fisher combiner seems to be the most appropriate $p$-value combiner for aggregating partial coherence results over frequencies and detecting edges across networks.

Looking at the different ways of implementing the Stepdown procedure, we can see that the regime using all frequencies has lower AFP rates compared to the regime using independent frequencies only, in all models. Additionally, the dependent stepdown scheme returns very similar AERs and AFN rates to the independent scheme. Therefore the dependent scheme outperforms the independent scheme, as it can achieve the same level of accuracy but with greater robustness to false positives. This also suggests that combiners that can handle correlated data points are better than their independent counterparts. This calls for the development of a Fisher combiner that is capable of handling dependent frequencies. As its performance in terms of AER and AFN rates is currently superior to those of other combiners, it could be improved even further when adapted to work across all data points, and help reduce its relatively high AFP rate. Some initial work
Fig. 1. AER (Eq. (7.2)), AFP and AFN (Eqs. (7.3)) of various p-value combiners tested at various levels, for five distinct 10-dimensional VAR(1) models with varying DoCs (x-axis, Definition 7.1). Dotted lines: independent frequencies combined. Solid lines: all frequencies combined.

7.2. Effect of group heterogeneity

The simulation study depicted in the previous section is repeated here for “families” of time series, which all share the same underlying parametric model but have different noise processes. Recall the definition of the true spectral matrices $\mathbf{S}(f)$, detailed in Appendix B. Variations can be made from the true matrices $\mathbf{S}(f)$ by inserting random deviations in the noise covariance matrix $\mathbf{\Sigma}_e$:

$$
\mathbf{\Sigma}_e = \{\Sigma_{e,kj}\}_{k,j=1}^p,
\Sigma_{e,kj} = \Sigma_{e,kj} + \mu_{k,j},
$$

(7.4)

$\mu_{k,j} \sim \text{i.i.d } U(-0.1,0.1)$.

Remark 8. The matrix $\mathbf{\Sigma}_e$ is thus transformed by a perturbation matrix $\mu$ such that: $\Sigma'_e = \Sigma_e + \mu$, $\mu = \mu'$ and $\mathbb{E}(\mu) = 0$.

We create $n_{\text{ind}} = 10$ perturbation matrices $\mu_1, \ldots, \mu_{10}$ and associate them to one of the $n_{\text{ind}}$ families derived from the 10-dimensional VAR(1) model #2, detailed in Table 1. In the same manner as in Section 7.1, $m = 500$ copies are created for each of the $n_{\text{ind}}$ families of the VAR(1) model. The results for each partial
coherence test (Eq. (3.2)) are then aggregated using the Fisher combiner of Section 4.3, as it produced better results in terms of AER in the previous section. We then evaluate the ability of the Stepdown procedure and of the Benjamini & Hochberg (B&H) procedure (Section 6) to retrieve the true connectivity network after combining p-values from all 10 families, with α = 0.05. Results are displayed in Fig. 3 in terms of AER (Eq. (7.2)), AFP rate and AFN rate (Eq. (7.3)) as a function of the threshold \( \theta \in [0, 1] \), depicted in Eq. (6.2).

From Fig. 3, we can see that the AFP rate drops and the AFN rate rises as a function of \( \theta \). This is expected, as a larger \( \theta \) is synonymous with stricter requirements in terms of edge detection across all subjects. Beyond \( \theta = 0.5 \), we can see that the AFP rate drops to zero, meaning that no false edge detection occurs. This suggests that no additional FDR correction is needed on group-wide graphs produced with the Stepdown or B&H conjunction protocols, for any \( \theta > 0.5 \).

In both the Stepdown and B&H procedure, an optimum can be achieved in terms of AER for thresholds \( \theta \in [0.3, 0.6] \). This understandably represents a balance between a low AFP rate and an AFN rate that is not excessively high, and also gives a reasonable set of choices for \( \theta \)s for the making of coloured graphs in Section 8.2.

8. Application to data set

In this section the results exposed in Section 6 are applied to the data set detailed below, in Section 8.1.
8.1. Data set

We focus our attention on the data set discussed in Medkour et al. (2010): EEGs are taken from 34 mental health patients (all male, aged between 20 and 60 years), in resting-state and with eyes closed, from the Serbsky Institute in Moscow. Of these patients, 15 were diagnosed with positive syndrome schizophrenia, 19 were diagnosed with negative syndrome Schizophrenia. These two conditions are quite distinct, notably in terms of their respective clinical symptoms. Hence, a motivation for this data collection is to understand whether these conditions can also be reliably identified using neurological imaging. Notably, we are interested in establishing whether it is possible to discriminate between positive and negative schizophrenia with graph theory techniques.

In addition, a control group of 24 healthy subjects of similar age to the patients is introduced to the study. We denote the positive syndrome group as “Positive”, the negative syndrome group as “Negative” and the control group as “Control”.

All subjects gave written informed consent before taking part in this investigation. Ethical approval came from the local Moscow Ethics Committee and in compliance with national legislation and the Declaration of Helsinki.

Previous research has found consistent differences in brain activity of schizophrenia patients compared to controls at low frequencies (Di Lorenzo et al., 2015; Itoh et al., 2011), especially in unmedicated patients (Mientus et al., 2002). Thus, the Delta band, defined over the frequency range [0.5, 4] Hz, is of high scientific interest in the context of this study.

Additionally, there exists strong neural oscillations, known as alpha rhythm, that create a dominant spectral line at around 10 Hz (Dawson and Fischer, 1994) which can interfere with the results. The alpha rhythm – otherwise known as alpha wave – evolves in the range of [8, 12] Hz and is caused by the electrical activity of the thalamic pacemaker cells in the occipital lobes (a.k.a in channels O1 and O2). Depending on the spectral window used in the estimation method, there can be transfers of power into other bands, including the delta band. It is important to get rid of this interference, as it could cause noticeable side-lobe leakage and significant estimation errors when analysing causal activity. Therefore we focus on the Delta band by means of a Low-Pass filter, as detailed below.

The EEGs in this data have been recorded once from distinct patients, under resting conditions and with eyes closed, meaning they could not have been any interactions between subjects during the experiment. On this basis, we expect to see no correlation between any of the recordings, and assume that they are independent of each other.

The EEGs are recorded for 30 sec. at a sampling rate of 100 Hz ($\Delta t = 0.01 \text{s}$, Nyquist frequency $f_N = \frac{1}{2\Delta t} = 50 \text{Hz}$) from $p = 10$ scalp sites, named according to the 10–20 system, F3, F4, C3, C4, T3, T4, P3, P4, O1 and O2, and referenced to linked ears with a band-pass filter of 0.5–45 Hz. The signals are then put through a 6 Hz Low-Pass Butterworth filter to remove leakage from the alpha rhythm. The signal is then downsampled by 5, in order to bring the Nyquist frequency to $f_N = 10 \text{Hz}$, so as to eliminate the zero-valued regions of the spectrum. The effects of this pre-processing are illustrated on the power spectrum of one channel from one patient in Fig. 4. The Low-Pass filter adequately manages to remove the alpha rhythm spike at around 10 Hz, while the down sampling adjusts the
spectrum over the zero-valued ranges. We can also see that most of the power in the pre-processed spectrum is constrained to the Delta band, in [0.5, 4] Hz. Hence, we focus our analysis in this frequency band specifically.

The resulting signal is a multivariate time series \( \{X_{ij}\} \) with dimension \( p = 10 \) where \( j = 1, \ldots, 10 \) represents the re-labelling of scalp sites:

<table>
<thead>
<tr>
<th>F3</th>
<th>F4</th>
<th>C3</th>
<th>C4</th>
<th>T3</th>
<th>T4</th>
<th>P3</th>
<th>P4</th>
<th>O1</th>
<th>O2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

The data of each patient can be viewed as a 2nd order stationary vector time series \( \{X_t\} \in \mathbb{R}^p, t \in \{0, \ldots, T-1\} \), as the EEG are recorded under resting-state and eyes-closed conditions.

For the data of each patient, the multitaper spectral estimate \( \hat{S}(f) \) (Eq. (2.1)) with \( K = 12 \) is computed, with a bandwidth \( B = 0.4221 \, \text{Hz} \) (Eq. (3.3)). A diagonal up-lift with a factor of \( \eta = 10^{-6} \) is applied in order to stabilise inversion.

The estimated partial coherences are then calculated from the inverted spectral matrix estimate \( \hat{S}^{-1}(f) \) at each discretized frequency \( f \). Some examples are displayed in Fig. 5, where we can see that some estimates can be quite large on average across the frequency band of interest. These estimates at each frequency need to be tested according to the protocol in Eq. (3.2) to ascertain significance. We then aggregate the results for each partial coherence using the Fisher \( p \)-value combiner of Section 4.3. Based on results of Section 7.2, we opt to use no additional FDR control, and keep the level \( \alpha \) at \( \alpha = 0.05 \). This produces one \( p \)-value for each patient in each group, which we wish to aggregate as detailed below.

### 8.2. Analysis

We investigate the ability of the Stepdown procedure and the Benjamini & Hochberg (B&H) protocol, as exposed in Section 6, to produce group-wide coloured graphs, using various thresholds \( \theta \), as described in Eq. (6.2):

\[
\theta \in \Theta = \{0.33, 0.5, 0.66\}.
\]

Results are displayed in Figs. 6 and 7. This protocol is applied to all the patients in all three groups, “Positive”, “Negative” and “Control”. Using these choices of thresholds \( \theta \in \Theta \), we can see the edges that are significant across three 33\%, 50\% and 66\% of patients in each group, respectively. For each group, a coloured graph is produced using the thresholds in \( \Theta \). The estimated graphs can be compared with one another at each threshold \( \theta \in \Theta \) using the Hamming distance detailed in Eq. (7.1).

Additionally, the independence between subjects in each group can be exploited to perform a bootstrap on the subjects time series. Denote \( N_g = 5000 \) to be the number of bootstraps copies. For each of the three groups \( g \leq N_g \) and each bootstrap copy \( n_b \leq N_g \), we perform the following:

1. Sample the time series of \( N_g \) subjects in group \( g \), selected at random with replacement from the set \( \{1, \ldots, N_g\} \), which we denote as \( \{b_{i1}^g, \ldots, b_{in}^g\} \).
2. Compute a group-specific graphs \( \hat{C}_g \) from \( \{b_{i1}^g, \ldots, b_{in}^g\} \) across a wider range of thresholds \( \theta \in [0, 1] \), using the method detailed in Sections 3.4–6.

The results are averaged across all bootstraps \( N_g \) and summarised in terms of expected Hamming distances between groups as a function of \( \theta \) in Fig. 8. Performing this bootstrapping procedure is an effective way to assess whether there is a true difference in connectivity between groups, and what is the sensitivity of the Stepdown and B&H procedures to the choice of \( \theta \).

In Figs. 6 and 7, we can see clear distinctions between each group of patients, for all values of \( \theta \in \Theta \). This holds across both combining methods, which return similar profiles for each group. Across both conjunction techniques, there are visible differences between the graphical models of the “Negative” and those of the two other groups. The group graph for the negative syndrome patient is significantly more connected in comparison with the other two groups. While the “Positive” group also differs form the controls, it does less so. The distance is greatest between the groups “Negative” and “Control” for any value of \( \theta \geq 0.15 \), suggesting that negative syndrome schizophrenia impacts the functional connectivity of the brain in a way that is more profound than the positive syndrome type. This is consistent with clinical observations made in patients: negative syndrome schizophrenia is considered to be a more severe form of the illness, as patients suffering from it see their ability to function and live independently greatly diminished, and have worse prognosis. Typical symptoms include reductions in speech, emotional withdrawal and apathy (Peralta and Cuesta, 1994). Positive syndrome patients on the other hand tend to experience delusions and/or hallucinations for instance, but are more likely to function better than their negative syndrome counterparts.

Fig. 8 shows the evolution of the expected hamming distance between all three groups as a function of \( \theta \in [0, 1] \). We can see that none of the groups have any hamming distance to each other for very high values of \( \theta \), suggesting that such thresholds return empty
Fig. 5. Estimated partial coherences, averaged across [0.5, 4] Hz, for three subjects chosen at random from the data set.

Fig. 6. All groups, Stepdown.

Fig. 7. All groups, B&H.

Coloured graphical models for the groups “Positive”, “Negative” and “Control” of Section 8.1, computed with the thresholds $\theta \in \Theta$ (Dotted edges: $\theta = 0.33$, Dashed: $\theta = 0.5$, solid: $\theta = 0.66$). Top: Stepdown procedure, as detailed in Eq. (4.1); Bottom: B&H procedure, as in Eq. (6.1).

Fig. 8. Bootstrapped average of the Hamming distance computed for group ‘P’ and ‘C’ (dotted line) and ‘N’ and ‘C’ (solid line) as a function of threshold $\theta$, after the significant proportions $\{P_{\delta}\}$ were computed by a Stepdown procedure (left) and B&H protocol (right).

graphs for each groups. Divergences between groups of patients become more apparent as the threshold $\theta$ is lowered, showing maximal distance between graphs for values of $\theta$ in [0.1, 0.3]. This observation is consistent across both protocols, suggesting that they can return agreeable results when applied to the same data set.

The value of $\theta$ directly impacts the Hamming distance between groups, and an interesting endeavour would be to select values of $\theta$
such that the resulting graphs are most distinct from each other. For the purpose of analysing the data of Section 8.1, this may not be the most relevant approach; we are more concerned with retrieving the true functional connectivity graphs of each mental health patients group, as opposed to maximising the distance hamming distances between these graphs.

Groups of subjects can also be compared using subject-specific graphs, and Hamming distances measured between individuals within and between groups. Using a bootstrap on the subjects time series, we can compute the means and empirical distributions for the intra-groups and inter-groups subject-specific Hamming distances, in Fig. 9 and Table 2. While the subject-to-subject distance between groups is strictly non-negative for any two groups, the expectation and variability of intra group distances are also non-negligibly large for all groups, rendering any form of comparison non-viable. On the other hand, the results of Table 2 highlights the substantial heterogeneity of subjects within each group. We can see more distances between the graphs within the “Negative” group compared to the two other groups, for instance. This is reflected in Figs. 6 and 7, where there are more edges of various colours in the graphical models of the “Negative” group. This indicates stronger heterogeneity in this group of patient. This can be explained by the diverse nature of Schizophrenia: clinically, the heterogeneity of this condition is widely recognised, with at least five sub-types being documented across three more general syndromes – “Positive”, “Negative” and “Disorganised” (Heinrichs and Awad, 1993).

**Remark 9.** While no FDR control is needed for the derivation of the coloured graphs of Figs. 6 and 7, it is necessary to use a FDR adjustment for the analysis of the subject-specific graphs examined in Fig. 9 and Table 2. In this case, due to the low dimensionality of the data ($p = 10$ dimensions) we opt for the regular Bonferroni correction of Eq. (5.2).

Using conjunction analysis on coloured graphical models, significant differences can be established between the various groups of subjects in the data set of Section 8.1. This supports the idea that EEG data can be used for the purpose of identifying various conditions. It remains to be seen whether or not, and how, this type of analysis can scale to neurological recordings of larger dimensions. The samples used in this study were collected from a $p = 10$ dimensional EEG set, and the results derived from it were obtained using methodologies suitable for low-dimensional data. It would be worthwhile to apply those graphical modelling methodologies to larger data – collected from larger EEG recordings for instance – to see whether similar results can be found.

### 9. Discussion

In this paper, the use of various p-value combiners have been demonstrated for the purpose of estimating graphical models for time series analysed in the frequency domain. This process is applied to EEG data collected from groups of psychiatric patients with distinct conditions, and split into two aspects.

First, estimates of pair-wise connectivity – the partial coherences – are tested for their significance at all frequencies, and p-values are aggregated across a suitable set of frequencies in the range of interest. In this paper the Fisher combiner was identified as the preferred choice in terms of average error rate, despite its high false positive rate and even when compared with the Stepdown procedure.

For any of the p-value combiners reviewed in this paper, the false positive rate can be effectively controlled across subject-specific networks, using a Bonferroni-type adjustment on the level $\alpha$ from Section 5, even on combiners with high False Positive Rates such as the Fisher combiner. This works well on data of relatively low dimensions, but quickly becomes too restrictive with higher dimensional data. An alternative based on Benjamini & Hochberg's protocol has been proposed, and shown to improve on the Bonferroni type adjustment in terms of False Negative Rate on higher dimensional data. However the improvement it achieves is moderate if the data becomes very high dimensional.

At this stage, we have results for each patient in the dataset, but need to derive group-wide graphs. As a next step, subject-specific results are aggregated to produce group-wide
results using techniques borrowed from conjunction analysis. Two thresholding methods are put forward to perform this step, which produce results that converge as the threshold is raised and efficiently remove the need for FDR control beforehand. These are applied to an EEG data set of unmedicated schizophrenia patients, separated into groups of positive/negative syndrome types and healthy controls. Distinct group-wide models are produced, in which the functional connectivity differences between positive/negative syndrome schizophrenia patients and healthy individuals are highlighted. Various levels of $\theta$ can be used to produce coloured graphs, showing how prevalent each connection is in each group. The coloured graphical model system is an efficient way of representing neurological data collected on groups of subjects, as it manages to show the level of heterogeneity present in each group of data. On the other hand, it does not give any indication as to which threshold level $\theta$ is optimal. It would be of interest to further research methods to aggregate results subjects-wide.

The data set investigated in this paper was of relatively low dimensions, with data recorded once from ten channels only. It would be highly interesting to apply the methods proposed in this paper – notably those of Section 6 on combining group-wide results – to higher dimensional data, such as those issued from larger EEGs, MEGs or fMRI.

This paper advocates the use of $p$-value combiners for graphical modelling in the frequency domain, and also raises questions that merit further research. Empirical evidence from simulation studies show that combiners that are able to deal with dependent data clearly outperform their counterparts using independent data in terms of false positive rate, but currently dependent versions of the Fisher $p$-value combiner do not exist in a form that is analytically satisfactory. The topic of developing combiners for dependent data has been covered in applications to Bio-statistics (Kost and McDermott, 2002; Chen et al., 2014; Li et al., 2014), but most of these approaches are analytically limited. Hence, there still exist a real need for $p$-value combiners that can effectively operate on dependent data, and especially on spectral data.

Acknowledgements

The author thanks their supervisor, Prof. Andrew T. Walden, as well as Dr Paul Ginzberg, for their comments and suggestions. The author’s work is supported by EPSRC (UK). The author is also thankful for the comments and suggestions made by the reviewers – notably for the formulation in Eq. (7.4) – which helped improve the paper substantially.

Appendix A. VAR(1) models

The VAR(1) times series models used in the simulation studies have been derived using a random parametric model generator adapted from (Wolstenholme and Walden, 2015, Appendix A): the algorithm produces a matrix $\Phi$ populated with standard normal random variable for every entry $(i,j)$ such as $(i+j)_{\text{mod}(k)} = 1$, where $k$ is a sparsity constant, and the eigenvalues of $\Phi$ are adjusted so as to ensure that the resulting VAR(1) model is stationary. While originally, the sparsity constant $k$ is kept fixed, here we make it vary over the range $k \in \{2, 4, 6, 8, 10\}$ in order to get VAR(1) models with various $\text{DoFs}$ (Definition 7.1). A lower sparsity constant $k$ returns a more highly connected VAR(1) model. In order to obtain the same VAR(1) models across simulations, we fix the random generator starting seed at seed = 2 in MatLab:

\[
\text{seed} = 2; \% \text{any strictly positive integer}
\]
\[
\text{rng(seed)}
\]

Appendix B. True spectra of VAR(1) models

The true spectral matrices $S(f)$ associated with a VAR(1) model defined as follows:

\[
X_t = \Phi X_{t-1} + \epsilon_t, \\
\epsilon_t \sim \text{i.i.d. } \mathcal{N}(0, \Sigma_e),
\]

have a closed-form known expression for all $f$ in range, which solely relies on the transfer matrix $\Phi$ and the noise covariance matrix $\Sigma_e$ (Percival and Walden, 1993):

\[
S(f) = \left( I - \sum_{i=1}^{T} \Phi^i \right)^{-1} S_{\text{noise}},
\]

Unless explicitly stated otherwise, the noise covariance $\Sigma_e$ is equal to an identity matrix: $\Sigma_e = I_p$, meaning that the noise process $\epsilon_t$ is white noise.

References


