Interpreting Null Models of Resting-State Functional MRI

Raphaël Liégeois[*,a,b], B. T. Thomas Yeo[c,d] and Dimitri Van De Ville[a,b]

[a] *Institute of Bioengineering, Center for Neuroprosthetics, École Polytechnique Fédérale de Lausanne, Switzerland*
[b] *Department of Radiology and Medical Informatics, University of Geneva, Switzerland*
[c] *Department of Electrical and Computer Engineering, National University of Singapore, Singapore*
[d] *N.1 Institute for Health & Institute for Digital Medicine (WisDM), National University of Singapore, Singapore*

Abstract

Null models are necessary for assessing whether a dataset exhibits non-trivial statistical properties. These models have recently gained interest in the neuroimaging community as means to explore dynamic properties of functional Magnetic Resonance Imaging (fMRI) time series. Interpretation of null-model testing in this context may not be straightforward because (i) null hypotheses associated to different null models are sometimes unclear and (ii) fMRI metrics might be 'trivial', i.e. preserved under the null hypothesis, and still be useful in neuroimaging applications. In this commentary, we review several commonly used null models of fMRI time series and discuss the interpretation of the corresponding tests. We argue that, while null-model testing allows for a better characterization of the statistical properties of fMRI time series and associated metrics, it should not be considered as a mandatory validation step to assess their relevance in neuroimaging applications.

Keywords: Null Models, Surrogate Data, Fourier Phase Randomization, Autoregressive Models.

Introduction

Functional interactions in the resting human brain are organized in complex spatio-temporal patterns (Greicius et al., 2003; Smith et al., 2013). The classical way of characterizing these patterns is within *static* modelling frameworks, e.g., evaluating functional connectivity (FC) from the correlation between whole-run functional Magnetic Resonance Imaging (fMRI) time series across different brain regions (Biswal et al., 1995; Zalesky et al., 2010; Margulies et al., 2016). In contrast, recent work has brought out promises of considering *dynamic* frameworks, such as sliding window

*To whom correspondence may be addressed.
Email: Raphael.Liegeois@epfl.ch

methods (Sakoğlu et al., 2010; Handwerker et al., 2012; Allen et al., 2014), to account for temporal variability of connectivity patterns (see Preti et al. (2017); Lurie et al. (2020) for recent reviews). Such dynamic FC frameworks raise questions like *'Are transition probability patterns between brain states different in the original data and in surrogate data with matched spatial covariance structure?'* or *'Are original fluctuations of a metric only due to autocorrelation of fMRI time series or are they revealing the presence of more complex (e.g., nonlinear) functional interactions?'*. These questions can be addressed using null-model testing that consists in comparing the original data vs. random data preserving a subset of original statistical properties, a.k.a. surrogate data (e.g., Theiler et al., 1992).

Null-model testing has been widely used in dynamic FC studies with apparently contradictory outcomes: while most studies concluded the rejection of the null hypothesis (Chang and Glover, 2010; Handwerker et al., 2012; Zalesky et al., 2014), others reported difficulties in rejecting the null hypothesis (Hindriks et al., 2016; Laumann et al., 2016; Liégeois et al., 2017). This observation is the starting point of our commentary. We first recall basics of null-model testing and provide a detailed description of popular null models of fMRI time series such as phase randomization and the autoregressive null model. As we will see, these null-model methods are related but test for distinct null hypotheses that need to be carefully understood in order to interpret testing outcomes. Then, we argue that a statistical metric preserved under a null hypothesis can still be useful in neuroimaging applications, thereby making null-model testing of limited relevance in this context. Finally, we discuss a few concrete examples and provide recommendations on the use and interpretation of null-model testing in dynamic FC studies.

Basics of null-model testing

We consider a time-series dataset $\mathbf{D}$ with dimensions $N \times T$ where N is the number of variables and T is the number of time points, and we denote $\mathbf{d}_t$ as the $N \times 1$ vector encoding values of the N variables at time t. Without loss of generality, we further assume that each of the N time courses composing $\mathbf{D}$ are centered and of unit variance. The autocorrelation sequence of $\mathbf{D}$ is defined as

the following sequence of $N \times N$ matrices:

$$\mathbf{R}_l = \frac{1}{T}\sum_{t=1}^{T} \mathbf{d}_t \mathbf{d}'_{t+l} \quad \text{for} \quad 0 \leq l \leq T-1, \tag{1}$$

where $'$ denotes transpose. Diagonal elements of $\mathbf{R}_l$ encode autocorrelation within individual time courses, while off-diagonal terms of $\mathbf{R}_l$ encode autocorrelation between pairs of time courses. Note that $\mathbf{R}_0$ encodes the classical correlation of $\mathbf{D}$ (or, equivalently, its covariance since time courses are assumed to be of unit variance) and we refer to it as Σ hereafter.

The goal of null-model testing is to gain insights into the statistical nature of $\mathbf{D}$ by testing the validity of a null hypothesis $\mathcal{H}$ to represent $\mathbf{D}$. Null-model testing frameworks typically consist of the three following steps.

1. Generate data preserving a subset of original statistical properties of $\mathbf{D}$, but that are otherwise random. This data is referred to as *surrogate* data and we denote it $\mathbf{D}_s$. Importantly, $\mathbf{D}_s$ comprises several data instances and the method used to generate $\mathbf{D}_s$ defines the null hypothesis $\mathcal{H}$.
2. Compute a metric m in both the original and surrogate datasets. The metric value in the original dataset is denoted as m_o, and the metric values computed from each surrogate data instance in $\mathbf{D}_s$ define a null distribution denoted $\mathcal{M}_s$.
3. Compare m_o and the null distribution $\mathcal{M}_s$. If m_o is different from $\mathcal{M}_s$, the null hypothesis $\mathcal{H}$ is not sufficient to fully describe $\mathbf{D}$. If m_o is not different from $\mathcal{M}_s$, the null hypothesis $\mathcal{H}$ cannot be rejected.

A toy example

Consider the case of a univariate dataset $\mathbf{D}$ with 50 time points, i.e., $N = 1$ and $T = 50$, shown in Figure 1A. To better characterize $\mathbf{D}$, let us apply two null-model approaches.

First, we consider the null hypothesis, denoted $\mathcal{H}_G$, that $\mathbf{D}$ can be represented by independent and identically distributed (i.i.d.) Gaussian random variables. Step 1 in the above procedure consists in producing surrogate data $\mathbf{D}_s$ from an i.i.d. Gaussian distribution with length, mean (μ) and standard deviation (σ) matched to the length, mean and standard deviation of $\mathbf{D}$ (Theiler

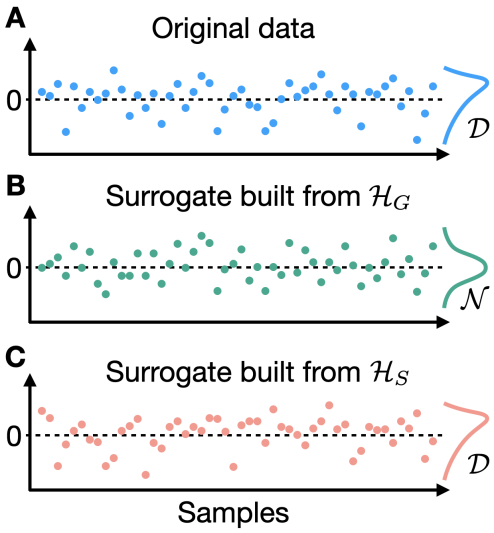

Figure 1: Comparing original and surrogate data. (A) Original data containing 50 time points with sample distribution $\mathcal{D}$. (B) One instance of surrogate data generated using a Gaussian univariate null hypothesis ($\mathcal{H}_G$). The corresponding (Gaussian) sample distribution is denoted $\mathcal{N}$ with mean and standard deviation matched to the original data. (C) One instance of surrogate data generated by shuffling original data. The null hypothesis corresponding to this procedure is $\mathcal{H}_S$ and the corresponding sample distribution is the same as the original data ($\mathcal{D}$).

et al., 1992). An instance obtained following this procedure is shown in Figure 1B. Then, in Step 2 one needs to compute a metric m in the original and surrogate data that is not directly defined by μ and σ as these quantities are by construction equal in $\mathbf{D}_s$ and $\mathbf{D}$. One could for example compare the number of time points above a given threshold in both the original and all surrogate datasets, leading to m_o and $\mathcal{M}_s$. Finally, in Step 3 values of m_o and $\mathcal{M}_s$ are compared in order to evaluate validity of the null hypothesis. In this case the null hypothesis, denoted $\mathcal{H}_G$, is straightforward and could be expressed as:

$$\mathcal{H}_G : d_t \sim \mathcal{N}(\mu, \sigma), \ \forall t \in [1, \ldots, 50]. \tag{2}$$

We will see in the next null-model framework that the null hypothesis defined by Step 1 is not always easy to identify. Then, the interpretation of the comparison performed in Step 3 depends on the null hypothesis: finding that m_o is different from $\mathcal{M}_s$ suggests that $\mathbf{D}$ contains statistical information beyond μ and σ, whereas finding that m_o is not different from $\mathcal{M}_s$ indicates that $\mathcal{H}_G$ cannot be rejected.

The second null-model framework we consider consists in generating surrogate data of Step 1

by shuffling time points of $\mathbf{D}$, as illustrated in Figure 1C (Scheinkman and LeBaron, 1989). The corresponding null hypothesis, denoted $\mathcal{H}_S$, is not explicitly defined by this procedure and could be expressed as:

$$\mathcal{H}_S : d_t \sim \mathcal{D}, \ \forall t \in [1, \ldots, 50], \tag{3}$$

where $\mathcal{D}$ is the sample distribution of $\mathbf{D}$ (Figure 1A). As such, $\mathbf{D}_s$ preserves the mean and standard deviation of $\mathbf{D}$ but also all other statistics defined by $\mathcal{D}$ such as kurtosis and skewness. The metric computed at Step 2 needs to be chosen accordingly: for example, comparing the number of time points above a given threshold in $\mathbf{D}$ and $\mathbf{D}_s$ would in this case be meaningless since this quantity is by construction equal in the original and surrogate datasets. An appropriate metric m could be, e.g., the average autocorrelation in $\mathbf{D}$ and $\mathbf{D}_s$. Finally, comparison of m_o and $\mathcal{M}_s$ in Step 3 provides information on validity of $\mathcal{H}_S$: finding that m_o is different from $\mathcal{M}_s$ shows there is autocorrelation in $\mathbf{D}$ whereas finding that m_o is not different from $\mathcal{M}_s$ suggests there is no temporal ordering information in $\mathbf{D}$.

Beyond illustrating main steps of null-model testing frameworks, several important observations can be made from this simple toy example. First, the null hypothesis associated to the surrogate data generation procedure (Step 1) is not always straightforward to identify. Second, it is crucial to clearly identify this null hypothesis in order to (i) define a relevant metric m to be compared in original and surrogate datasets (Step 2) and (ii) properly interpret the result of this comparison (Step 3). Third, null hypothesis rejection does not have a unique interpretation as it can be due to violation of any property preserved by the null. For example, rejecting $\mathcal{H}_G$ could be due to non-Gaussianity, non-stationarity, the presence of autocorrelation in $\mathbf{D}$, or a combination of these. A careful choice of the metric m in Step 2 might in certain cases allow to narrow down causes of null-model rejection, i.e., in this case determining which statistical property among non-Gaussianity, non-stationarity, or autocorrelation, led to the rejection of $\mathcal{H}_G$. The description of the different metrics types is beyond the scope of this paper and we refer the reader to the excellent review by Lancaster et al. (2018) for a detailed presentation of these metrics and the statistical properties they capture. Finally, note that not rejecting the null hypothesis at Step 3 using a given metric m

does not imply that the null model is sufficient to fully describe $\mathbf{D}$.

Null models of fMRI time series

The two null models presented in the above toy example are not suited to explore the statistical properties of fMRI time series because the latter are known to be autocorrelated, leading to an -uninformative- rejection of $\mathcal{H}_G$ and $\mathcal{H}_S$. We present two null-model frameworks that account for this autocorrelation while also preserving fMRI covariance structure: multivariate autoregressive null models and phase randomization (Theiler et al., 1992).

Multivariate autoregressive (AR) null models assume that $\mathbf{D}$ can be represented as an autoregressive model. We focus on the case of first-order AR (AR-1) models, i.e., the null hypothesis, denoted $\mathcal{H}_A$, is that data at time t can be expressed as a weighted sum of data at time $t-1$ plus Gaussian noise:

$$\mathcal{H}_A : \mathbf{d}_t = \mathbf{A} \cdot \mathbf{d}_{t-1} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{C}) \ \forall t, \tag{4}$$

where $\mathbf{A}$ is a model parameter of size $N \times N$ and ϵ_t is i.i.d. Gaussian noise. Each null sample in $\mathbf{D}_s$ is initialized by randomly selecting a time point in $\mathbf{D}$, and then generating $T-1$ new time points by successively applying Eq. (4), with $\mathbf{A}$ and $\mathbf{C}$ identified from the original data. It can be shown from the Yule-Walker equations that $\mathbf{D}$ and $\mathbf{D}_s$ have the same correlation (Σ) and first-order autocorrelation ($\mathbf{R}_1$, see Eq. (1)) structures (Yule, 1927; Walker, 1931). More details on statistical properties of AR null models and their application to fMRI time series are found in Liégeois et al. (2017).

Phase randomization (PR) is another method that has been widely used to explore properties of fMRI time series (e.g., Handwerker et al., 2012; Hindriks et al., 2016). In this framework, surrogate data of Step 1 is generated by performing Discrete Fourier Transform (DFT) of each time course, adding a uniformly distributed random phase to each frequency, and then performing the inverse DFT (Prichard and Theiler, 1994). Importantly, the random phases need to preserve Hermitian symmetry and are generated independently for each frequency, but they are the same across the N variables, i.e., brain regions or voxels in this case. From the Wiener-Khintchine theorem, it follows

that surrogate data generated by PR preserves the full autocorrelation structure of $\mathbf{D}$ defined in Eq. (1) (Wiener, 1930; Khintchine, 1934). As such, the corresponding null hypothesis is that $\mathbf{D}$ can be represented using a linear and Gaussian model:

$$\mathcal{H}_P : \mathbf{d}_t = \sum_{i=1}^{T} \mathbf{A}_i \cdot \mathbf{d}_{t-i} + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{C}) \; \forall t, \tag{5}$$

where $\mathbf{A}_i$ are model parameters of size $N \times N$ and ϵ_t is i.i.d. Gaussian noise. It can be seen by comparing Eqs. (4) and (5) that while both the AR-1 null and PR assume a linear and Gaussian representation of $\mathbf{D}$, PR preserves more statistical information from the original data. Indeed, the AR-1 null preserves Σ and $\mathbf{R}_1$ whereas PR preserves Σ and the whole autocorrelation sequence $\mathbf{R}_1, \ldots, \mathbf{R}_T$ defined in Eq. (1). We also note that while the model parameter $\mathbf{A}$ can be identified in Eq. (4) provided at least $N+1$ time points are available in $\mathbf{D}$ (i.e., $T > N$), PR model parameters $\mathbf{A}_i$ for $i = 1, \ldots, T$ in Eq. (5) cannot be explicitly identified from the autocorrelation sequence of $\mathbf{D}$ (Stoica and Moses, 2005). Despite these practical differences, both approaches were found to yield similar conclusions about the nature of fMRI time series which suggests that $\mathbf{R}_1$ captures a significant proportion of fMRI autocorrelation structure (Liégeois et al., 2017).

The main properties of the four null models introduced in this paper are summarized in Table 1. The first two frameworks corresponding to $\mathcal{H}_G$ and $\mathcal{H}_S$ do not preserve any temporal structure of time series data, thereby leading to rejection of the corresponding null hypotheses when applied to fMRI time series. We have included them in order to provide a complete description of the links between popular null-model frameworks, while also motivating the use of PR and the multivariate AR null in this context.

Related models and methods

PR and the AR null are most commonly used to explore properties of measures of interest computed from fMRI time series, but other methods have been considered. For example, Chang and Glover (2010) used a variation of the AR null that, instead of identifying a single multivariate AR model for the whole N-variate system of fMRI time series as described in Eq. (4), generates AR surrogates for each pair of variables separately. As a result, bivariate surrogate time series are

Table 1: Main properties of the four null-model frameworks presented in this commentary. For each framework, the corresponding null hypotheses and statistical properties of $\mathbf{D}$ being preserved in $\mathbf{D}_s$ are reported. A non-exhaustive list of valid measures m, i.e., measures that are not entirely defined by the statistical properties shared by $\mathbf{D}$ and $\mathbf{D}_s$, to be computed in Step 2 are also provided. $\mathcal{N}$: i.i.d. Gaussian distribution - $\mathcal{D}$: sample distribution of $\mathbf{D}$ - κ: kurtosis - Σ: correlation matrix - $\mathbf{R}_p$: p^{th} order autocorrelation matrix as defined in Eq. (1).

	Gaussian and i.i.d.	Data shuffling	AR-1 Randomization	Phase Randomization
Null hypothesis	$\mathcal{H}_G$: $\mathbf{d}_t = \epsilon_t, \epsilon_t \sim \mathcal{N}$	$\mathcal{H}_S$: $\mathbf{d}_t = \epsilon_t, \epsilon_t \sim \mathcal{D}$	$\mathcal{H}_A$: $\mathbf{d}_t = \mathbf{A} \cdot \mathbf{d}_{t-1} + \epsilon_t$	$\mathcal{H}_P$: $\mathbf{d}_t = \sum_{i=1}^{T} \mathbf{A}_i \cdot \mathbf{d}_{t-i} + \epsilon_t$
Properties preserved	Σ	$f(\mathcal{D}) : \Sigma$, κ, etc.	Σ, $\mathbf{R}_1$	Σ, $\mathbf{R}_1, \ldots, \mathbf{R}_T$
Examples of useful m	• Kurtosis (κ) • Nb. of points $>$ threshold • $\mathbf{R}_1$	• $\mathbf{R}_1$ • Measures of nonlinearity or nonstationarity	• $\mathbf{R}_p$ with $p > 1$ • Kurtosis (κ) • Measures of nonlinearity	• Measures of nonlinearity • State transition patterns

independent for each pair of variables and do not account for potential higher-order spatial interactions present in fMRI time series, thereby causing rejection of the corresponding null hypothesis (e.g., Figure 6 in Liégeois et al., 2017). Then, a variant of PR in which the random phase added to the DFT components of a given frequency is different for each variable has been introduced recently (Abrol et al., 2017). This framework, referred to as 'inconsistent' PR, does not preserve the correlation structure of $\mathbf{D}$ which causes rejection of the corresponding null hypothesis when applied to fMRI time series. Another recent framework produces surrogate data that preserves original correlation structure as well as the power spectral density of each individual time course, which by the Wiener-Khintchine theorem amounts to preserving the autocorrelation structure within each time course (Laumann et al., 2016). In other words, this framework preserves original Σ as well as the diagonal entries of the autocorrelation sequence $\mathbf{R}_1, \ldots, \mathbf{R}_T$ of $\mathbf{D}$ and could therefore be considered as an intermediate between the Gaussian i.i.d. null and PR in terms of statistical properties being preserved. Finally, other approaches factor in underlying anatomical information (Pirondini et al., 2016; Petrovic et al., 2020) or account for more complex statistical properties (e.g., Theiler et al., 1992; Breakspear et al., 2003; Van De Ville et al., 2004).

Statistically 'trivial' does not mean useless

When evaluating the relevance of a dynamic fMRI (or FC) metric, it has become common practice to require that this metric produces different results in original and surrogate data in order to ensure it does not reflect 'trivial' statistical data properties (see, e.g., Preti et al., 2017, and references therein). The word 'trivial' here refers to statistical properties preserved by the null-model framework applied to fMRI time series, usually multivariate AR or PR. While we agree these statistical properties are to some extent 'trivial' as $\mathcal{H}_A$ or $\mathcal{H}_P$ only preserve linear and Gaussian fMRI features, we argue that rejecting (or not rejecting) these null hypotheses is uninformative about the neurological or clinical use of the considered fMRI metric.

To this end, let us reformulate the interpretation of Step 3 testing outcome using the representation of Figure 2. The outer green circle represents the set of all statistical properties characterizing original data $\mathbf{D}$, i.e., all spatial, temporal, and spatiotemporal statistical moments of $\mathbf{D}$. The three inner circles represent statistical information preserved by three different null frameworks: Gaussian and i.i.d., AR-1, and PR. We did not include data shuffling for simplicity purposes as the corresponding null hypothesis is not nested within the other hypotheses, i.e., $\mathcal{H}_S$ contains strictly more information than $\mathcal{H}_G$ but is not contained in $\mathcal{H}_A$.

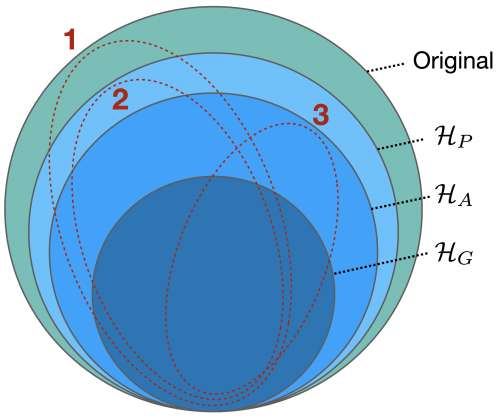

Figure 2: Nested structure of original statistical properties (outer green circle) being preserved by $\mathcal{H}_G$ (Gaussian i.i.d. null model, inner dark blue circle), $\mathcal{H}_A$ (AR-1 null, intermediate medium blue circle), and $\mathcal{H}_P$ (phase randomization, intermediate light blue circle). Red dotted ellipses represent different subsets of statistical information used in the practical examples at the end of the paper.

Rejecting a null hypothesis at Step 3 shows that the metric being considered exploits statistical properties *beyond* the ones preserved by the null hypothesis. For example, observing that an fMRI

metric is different in original and AR-1 surrogates suggests that this metric exploits statistical information beyond $\mathcal{H}_A$, i.e., beyond fMRI correlation (Σ) and first-order autocorrelation ($\mathbf{R}_1$) structures (e.g., this could be the case of a metric exploiting statistical information of curves 1 and 2 in Figure 2). Framed differently, the metric exploits some forms of nonlinearity, non-Gaussianity, non-stationarity, or higher-order autocorrelation of fMRI time courses. Importantly, exploiting fMRI statistical information beyond Σ and $\mathbf{R}_1$ is not informative about relevance of that metric in neuroimaging applications such as disease classification, fingerprinting, etc. Indeed, there is no reason for a metric outside of $\mathcal{H}_A$, e.g., a measure of nonlinearity such as kurtosis, to be *a priori* more relevant than a metric that lies within $\mathcal{H}_A$ -and therefore called 'trivial'- such as average first-order autocorrelation. In sum, the statistical information captured by fMRI or FC metrics and assessed using the null-model frameworks presented in this paper is unrelated to their clinical use which should be tested within a separate framework.

As an illustration, consider the case of fMRI time series correlation (Σ) that is classically used to evaluate FC. Utility of this metric in a wide range of neuroimaging applications no longer needs to be proven (e.g., Buckner et al., 2013, and references therein), and any metric derived from Σ such as partial correlations (Ryali et al., 2012) or graph metrics computed from Σ (Meunier et al., 2010; Mišić et al., 2016) and Σ^{-1} (Liégeois et al., 2020) would be found to be the same in original and surrogate data generated from $\mathcal{H}_G$, $\mathcal{H}_S$, $\mathcal{H}_A$ or $\mathcal{H}_P$ as all these null hypotheses preserve the original Σ (Table 1). This further illustrates that 'trivial' statistical information preserved within a null-model framework can be useful in neuroimaging applications. Note that metrics that are statistically equivalent and hence preserved under the same null-hypotheses, such as Σ^{-1} and Σ, can be of different usefulness; e.g., FC evaluated from *partial* correlations (derived from Σ^{-1}) was shown to outperform FC evaluated from correlations in several prediction tasks (Dadi et al., 2019). A similar rationale applies for measures derived from the autocorrelation sequence of fMRI time series that is preserved by $\mathcal{H}_A$ and $\mathcal{H}_P$, and was shown to provide insight into brain functional organization (Mitra et al., 2014) and its link with human behaviour (Liégeois et al., 2019).

A measure of model complexity

In the above we argue that null-model testing outcomes are not informative about relevance of a given metric in neuroimaging applications. This being said, these tests provide insight into the statistical complexity of the metric which could be useful in terms of model selection. For example, finding that a metric is shown to be equivalent in original and AR-1 surrogates suggests that it is exploiting statistical properties preserved by $\mathcal{H}_A$: Σ and $\mathbf{R}_1$. These two parameters can be represented by $N \cdot (3N-1)/2$ parameters (N^2 for $\mathbf{R}_1$ and $N \cdot (N-1)/2$ for Σ) which gives an upper bound on the parametric complexity of the metric. As for PR, the full autocorrelation sequence preserved by $\mathcal{H}_P$ is described by $\approx T \cdot N^2/2$ parameters[1]. This highlights that the notion of 'trivial' is relative: while PR only preserves linear and Gaussian properties, the parametric space defined by such models can get considerable when considering typical fMRI time series with $N \approx 100$ and $T \approx 1000$. We show in the first practical example hereunder how these consideration can be included to perform model selection of fMRI time series.

A null hypothesis can (almost) always be rejected

Rejecting a null hypothesis means that the corresponding null model is not sufficient to fully describe $\mathbf{D}$. This is correct, but should not be over-interpreted. As an illustration, consider the distribution of US women height obtained from the US Census Bureau[2] and represented by the grey bars in Figure 3.

Let us evaluate the relevance of a Gaussian representation of this distribution using null-model testing. Step 1 consists in identifying a Gaussian model from the data and generating surrogate data from it. The corresponding null hypothesis $\mathcal{H}_G$ is characterized by a mean μ (in this case $\mu = 64\ in$) and a standard deviation σ (in this case $\sigma = 2.6\ in$). Let us now define m used in Step 2 as the number of women taller than 80 inches. Under the Gaussian null, probability of finding a woman meeting this criterion is $p = 3.8 \cdot 10^{-10}$ which leads to rejection of the Gaussian null at Step

[1] The 1/2 factor comes from the fact that PR assumes periodic time series (Theiler et al., 1992). Therefore, $\mathbf{R}_p = R'_{T-p}$ for $0 \leq p \leq T$, and only half of the autocorrelation sequence can be described by independent parameters. The exact number of independent parameters depends on T being odd or even.

[2] https://www2.census.gov/library/publications/2010/compendia/statab/130ed/tables/11s0205.pdf

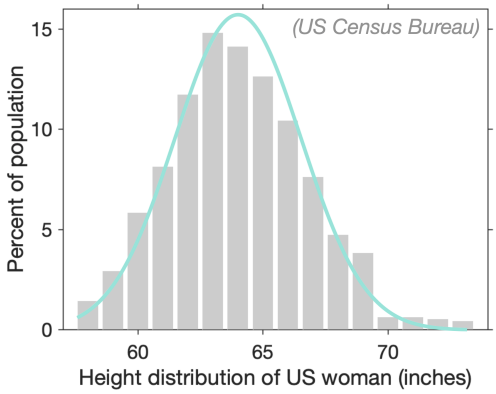

Figure 3: Data models are imperfect. Grey bars represent original distribution of height of US women aged 30-39 and the green line represents a Gaussian distribution matched to original data. A Gaussian model does not account for tall women (> 80 *in*), yet it provides a useful representation of US women height distribution.

3 since m_o is larger than the null distribution $\mathcal{M}_s$ (i.e., the number of women taller than 80 *in* is in reality non-zero). In other words, the Gaussian null is rejected because it does not fully capture properties of the data, in this case the proportion of extreme values of US women heights. Yet, it seems reasonable to consider that a Gaussian model of the data as represented by the green line in Figure 3 is informative about US women height distribution.

Two conclusions shall be drawn from this example. First, it seems virtually always possible to choose a metric m at Step 2 that leads to rejection of the null hypothesis by focusing on known peculiarities of the original dataset, e.g., the distribution of extreme values as illustrated above. Second, null hypothesis rejection does not imply that the corresponding model of the data is useless. Interpretation of this result should be performed in light of the more general model selection question, as illustrated in the first practical example hereunder.

Practical examples

We present three practical examples inspired from the dynamic FC literature. While we aimed to consider realistic applications and results, we did not actually perform the analyses as our goal is to discuss the interpretation of results typically encountered in null-model testing analyses of functional dynamics.

Example 1: Brain states dynamics

Let us consider a classical sliding window correlation (SWC) approach of fMRI time series followed by a clustering of FC connectivity matrices into so-called brain states. We use a clustering into five states and focus on the transition probability matrix that encodes transition probability between each pair of brain states (e.g., Allen et al., 2014). This FC dynamics summary metric, as well as related measures such as dwell time, were for example used to describe differences between control and schizophrenia groups (Damaraju et al., 2014; Du et al., 2016). Assume we compare the original transition probability matrix of one subject and the distribution of transition probability matrices computed in AR-1 surrogate data, as illustrated in Figure 4A, and find that three matrix entries are different in original and surrogate matrices. This result suggests that the transition probability matrix captures, at least partly, statistical information above and beyond the one preserved by the AR-1 null (e.g., curves 1 or 2 in Figure 2). More precisely, results of Figure 4A suggest that three entries of the transition probability matrix rely upon original statistical information beyond $\mathcal{H}_A$, while the 22 remaining entries out of 25 exploit information within $\mathcal{H}_A$.

Based on the considerations presented in the above sections, we make the following comments. First, similar to the case discussed in Figure 3, rejecting $\mathcal{H}_A$ does not disqualify the AR-1 representation of the data. This question has to be considered in a more general model selection framework balancing model fit (in this case, 22 out of 25 entries are reproduced by the AR-1 model) and model complexity. Then, while results show that transition probabilities from state 1 to 5, from state 2 to 1, and from state 5 to 3 reflect more complex (i.e., beyond $\mathcal{H}_A$) functional dynamics, this is uninformative about neurological relevance or clinical use of these matrix entries as compared to other matrix entries. Finally, considering five brain states corresponds to using $5N(N-1)/2$ parameters to summarize functional dynamics whereas the AR-1 representation uses $N(3N-1)/2$ parameters. This increased model complexity could be taken into account when discussing costs and benefits of brain state representations.

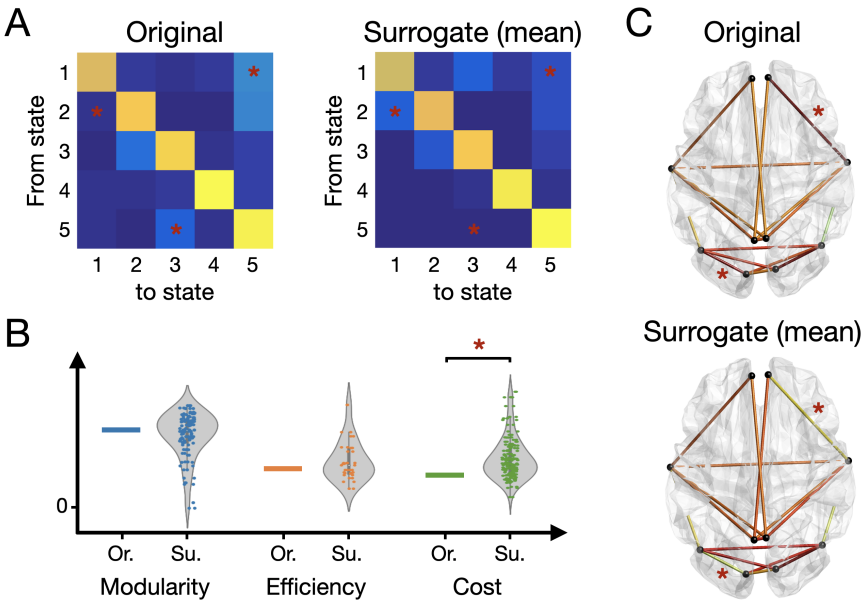

Figure 4: Illustrations of the three practical examples. Red stars denote significant differences between original and surrogate measures. (A) Transition probability matrices between five brain states computed in original (left) and surrogate (right) data. Warmer colors denote higher transition probabilities. Three matrix entries are found to be different between original and surrogate data: transitions probability from state 1 to 5, from state 2 to 1, and from state 5 to 3. (B) Three graph-theoretic measures are computed in original (Or.) and surrogate (Su.) data. All surrogate values are represented (i.e., not only the mean), and the corresponding violin plot is represented. One metric (cost) is found to be higher in surrogate than in original data. (C) Amplitude of sliding window correlation (SWC) fluctuations in a selection of 18 edges pertaining to the visual and default mode networks, in original (top) and surrogate (bottom) data. Two pairs of nodes are found to exhibit different SWC fluctuations in original and surrogate data.

Example 2: Graph metrics

Let us focus on graph metrics computed from brain state transition networks. These networks are computed by first applying a SWC approach, and then defining a network based on the transitions between successive FC states, as for example proposed in Ramirez-Mahaluf et al. (2020). Three graph metrics are considered: modularity, efficiency, and cost. These metrics are computed in both the original (Or.) and PR surrogate (Su.) data samples, as illustrated in Figure 4B. We assume no significant difference is found between original and surrogate data for modularity and efficiency, while the cost is found to be significantly higher in surrogate data as compared to original data. This suggests that modularity and efficiency reflect linear and Gaussian features of the original time series (e.g., curves 2 or 3 in Figure 2), whereas cost exploits at least partly nonlinear or non-Gaussian features (e.g., curve 1 in Figure 2). As in the previous example, this result is uninformative about the neuroscientific relevance or clinical use of the three metrics. Finally, note

that it would not be relevant to perform a null-model analysis on metrics derived from a functional graph defined by Σ (e.g., Hallquist and Hillary, 2019), since Σ is by construction preserved by both $\mathcal{H}_A$ and $\mathcal{H}_P$.

Example 3: Amplitude of SWC fluctuations

Amplitude of SWC fluctuations, formally evaluated from the variance of SWC time series, has also been used to describe functional dynamics (e.g., Liégeois et al., 2016; Hindriks et al., 2016). In this example we focus on the amplitude of SWC in a selection of 18 pairs of brain nodes (i.e., edges) located in the visual and default mode networks, computed both in original and AR-1 surrogate data (Figure 4C). Each edge reflects variance of SWC time series for the two corresponding nodes, with warmer colors denoting higher variance. Results presented in Figure 4C show that among the 18 edges considered in this analysis, no statistical difference between original and surrogate data is found for 16 of them, whereas 2 are found to be different. This suggests that fluctuations of SWC between the nodes corresponding to these two edges reflect, at least partly, nonlinear, non-Gaussian, or higher order linear functional interactions (e.g., curves 1 or 2 in Figure 2). In contrast, functional interactions represented by the 16 remaining edges is a priori capturing statistical information within $\mathcal{H}_A$ (e.g., curve 3 in Figure 2). Again, this result is uninformative about the neuroscientific relevance or clinical use of functional interactions represented by these two edges, as compared to the remaining 16 edges.

Conclusions and Recommendations

Null-model testing is widely used in dynamic FC studies. In this commentary we show that the choice of an appropriate null model, as well as the interpretation of the testing outcomes, requires caution. Concretely, we make the following suggestions on the use and interpretation of null-model testing of brain functional dynamics.

- We recommend using the multivariate autoregressive null or phase randomization to generate surrogate data of fMRI time series. It is important to be aware of the corresponding null hypotheses reported in Table 1 in order to interpret testing outcomes.

- Comparing a metric in original and surrogate data is uninformative about clinical or neurological utility of that metric. Therefore, null-model testing should *not* be considered as a mandatory step to validate a novel (dynamic) FC metric.

- Comparing a metric in original and surrogate data provides a better characterization of statistical information exploited by that metric. This can be used, e.g., to assess its parametric complexity.

While highlighting important limitations of null-model testing to explore functional dynamics, these considerations also draw promising avenues for defining new statistical markers of fMRI time series. Indeed, we have seen that statistical measures preserved by AR-1 or PR null models are a priori as relevant as more complex metrics to describe functional dynamics. In this context, fMRI time series' correlation has received a lot of attention as a marker of functional connectivity, but quite surprising is the fact that 'next circles' of statistical properties preserved by linear and Gaussian null models represented in Figure 2, i.e., their autocorrelation structure, has received limited attention so far.

Overall, this commentary provides the theoretical background to design and interpret null-model testing of brain functional time series. Beyond warning about potential misuse in dynamic FC studies, we highlight promises of this framework to identify novel metrics exploiting simple and interpretable features of functional dynamics.

Acknowledgements

RL acknowledges support by the Swiss National Centre of Competence in Research - Evolving Language (grant number 51NF40_180888).

References

Abrol, A., Damaraju, E., Miller, R.L., Stephen, J.M., Claus, E.D., Mayer, A.R., Calhoun, V.D., 2017. Replicability of time-varying connectivity patterns in large resting state fmri samples. Neuroimage 163, 160–176. doi:10.1016/j.neuroimage.2017.09.020.

Allen, E.A., Damaraju, E., Plis, S.M., Erhardt, E.B., Eichele, T., Calhoun, V.D., 2014. Tracking whole-brain connectivity dynamics in the resting state. Cerebral cortex 24, 663–676. doi:10.1093/cercor/bhs352.

Biswal, B., Yetkin, F.Z., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. Magnetic resonance in medicine 34, 537–541.

Breakspear, M., Brammer, M., Robinson, P.A., 2003. Construction of multivariate surrogate sets from nonlinear data using the wavelet transform. Physica D: Nonlinear Phenomena 182, 1 – 22. URL: http://www.sciencedirect.com/science/article/pii/S0167278903001362, doi:https://doi.org/10.1016/S0167-2789(03)00136-2.

Buckner, R.L., Krienen, F.M., Yeo, B.T., 2013. Opportunities and limitations of intrinsic functional connectivity mri. Nature neuroscience 16, 832–837.

Chang, C., Glover, G.H., 2010. Time–frequency dynamics of resting-state brain connectivity measured with fmri. Neuroimage 50, 81–98.

Dadi, K., Rahim, M., Abraham, A., Chyzhyk, D., Milham, M., Thirion, B., Varoquaux, G., 2019. Benchmarking functional connectome-based predictive models for resting-state fmri. NeuroImage 192, 115 – 134. doi:https://doi.org/10.1016/j.neuroimage.2019.02.062.

Damaraju, E., Allen, E., Belger, A., Ford, J., McEwen, S., Mathalon, D., Mueller, B., Pearlson, G., Potkin, S., Preda, A., et al., 2014. Dynamic functional connectivity analysis reveals transient states of dysconnectivity in schizophrenia. NeuroImage: Clinical 5, 298–308.

Du, Y., Pearlson, G.D., Yu, Q., He, H., Lin, D., Sui, J., Wu, L., Calhoun, V.D., 2016. Interaction among subsystems within default mode network diminished in schizophrenia patients: A dynamic connectivity approach. Schizophrenia research 170, 55–65.

Greicius, M.D., Krasnow, B., Reiss, A.L., Menon, V., 2003. Functional connectivity in the resting brain: a network analysis of the default mode hypothesis. Proceedings of the National Academy of Sciences 100, 253–258.

Hallquist, M.N., Hillary, F.G., 2019. Graph theory approaches to functional network organization in brain disorders: A critique for a brave new small-world. Network Neuroscience 3, 1–26.

Handwerker, D.A., Roopchansingh, V., Gonzalez-Castillo, J., Bandettini, P.A., 2012. Periodic changes in fmri connectivity. Neuroimage 63, 1712–1719.

Hindriks, R., Adhikari, M.H., Murayama, Y., Ganzetti, M., Mantini, D., Logothetis, N.K., Deco, G., 2016. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fmri? Neuroimage 127, 242–256.

Khintchine, A., 1934. Korrelationstheorie der stationaryren stochastischen prozesse. Mathematische Annalen 109.

Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V., Stefanovska, A., 2018. Surrogate data for hypothesis testing of physical systems. Physics Reports 748, 1–60. URL: https://www.sciencedirect.com/science/article/pii/S0370157318301340, doi:https://doi.org/10.1016/j.physrep.2018.06.001. surrogate data for hypothesis testing of physical systems.

Laumann, T.O., Snyder, A.Z., Mitra, A., Gordon, E.M., Gratton, C., Adeyemo, B., Gilmore, A.W., Nelson, S.M., Berg, J.J., Greene, D.J., et al., 2016. On the stability of bold fmri correlations. Cerebral Cortex , 1–14.

Liégeois, R., Laumann, T.O., Snyder, A.Z., Zhou, J., Yeo, B.T.T., 2017. Interpreting temporal fluctuations in resting-state functional connectivity mri. Neuroimage 163, 437–455. doi:10.1016/j.neuroimage.2017.09.012.

Liégeois, R., Li, J., Kong, R., Orban, C., Van De Ville, D., Ge, T., Sabuncu, M.R., Yeo, B.T.T., 2019. Resting brain dynamics at different timescales capture distinct aspects of human behavior. Nature Communications 10, 2317. doi:10.1038/s41467-019-10317-7.

Liégeois, R., Santos, A., Matta, V., Van De Ville, D., Sayed, A.H., 2020. Revisiting correlation-based functional connectivity and its relationship with structural connectivity. Network Neuroscience , 1–25URL: http://dx.doi.org/10.1162/netn_a_00166, doi:10.1162/netn_a_00166.

Liégeois, R., Ziegler, E., Phillips, C., Geurts, P., Gómez, F., Bahri, M.A., Yeo, B.T.T., Soddu, A., Vanhaudenhuyse, A., Laureys, S., Sepulchre, R., 2016. Cerebral functional connectivity periodically (de)synchronizes with anatomical constraints. Brain structure and function 221, 2985–2997. doi:10.1007/s00429-015-1083-y.

Lurie, D.J., Kessler, D., Bassett, D.S., Betzel, R.F., Breakspear, M., Kheilholz, S., Kucyi, A., Liégeois, R., Lindquist, M.A., McIntosh, A.R., et al., 2020. Questions and controversies in the study of time-varying functional connectivity in resting fmri. Network Neuroscience 4, 30–69.

Margulies, D.S., Ghosh, S.S., Goulas, A., Falkiewicz, M., Huntenburg, J.M., Langs, G., Bezgin, G., Eickhoff, S.B., Castellanos, F.X., Petrides, M., Jefferies, E., Smallwood, J., 2016. Situating the default-mode network along a

principal gradient of macroscale cortical organization. Proceedings of the National Academy of Sciences of the United States of America 113, 12574–12579. doi:10.1073/pnas.1608282113.

Meunier, D., Lambiotte, R., Bullmore, E.T., 2010. Modular and hierarchically modular organization of brain networks. Frontiers in Neuroscience 4, 200. doi:10.3389/fnins.2010.00200.

Mišić, B., Betzel, R.F., de Reus, M.A., van den Heuvel, M.P., Berman, M.G., McIntosh, A.R., Sporns, O., 2016. Network-level structure-function relationships in human neocortex. Cerebral Cortex 26, 3285–96. doi:10.1093/cercor/bhw089.

Mitra, A., Snyder, A.Z., Hacker, C.D., Raichle, M.E., 2014. Lag structure in resting-state fmri. Journal of neurophysiology 111, 2374–2391.

Petrovic, M., Liegeois, R., Bolton, T.A.W., Van De Ville, D., 2020. Community-aware graph signal processing: Modularity defines new ways of processing graph signals. IEEE Signal Processing Magazine 37, 150–159. doi:10.1109/MSP.2020.3018087.

Pirondini, E., Vybornova, A., Coscia, M., Ville, D.V.D., 2016. A spectral method for generating surrogate graph signals. IEEE Signal Processing Letters 23, 1275–1278. doi:10.1109/lsp.2016.2594072.

Preti, M.G., Bolton, T.A., Van De Ville, D., 2017. The dynamic functional connectome: State-of-the-art and perspectives. Neuroimage 160, 41–54. doi:10.1016/j.neuroimage.2016.12.061.

Prichard, D., Theiler, J., 1994. Generating surrogate data for time series with several simultaneously measured variables. Physical Review Letters 73, 951–954.

Ramirez-Mahaluf, J.P., Medel, V., Tepper, Á., Alliende, L.M., Sato, J.R., Ossandon, T., Crossley, N.A., 2020. Transitions between human functional brain networks reveal complex, cost-efficient and behaviorally-relevant temporal paths. Neuroimage 219, 117027. doi:10.1016/j.neuroimage.2020.117027.

Ryali, S., Chen, T., Supekar, K., Menon, V., 2012. Estimation of functional connectivity in fmri data using stability selection-based sparse partial correlation with elastic net penalty. Neuroimage 59, 3852–3861.

Sakoğlu, Ü., Pearlson, G.D., Kiehl, K.A., Wang, Y.M., Michael, A.M., Calhoun, V.D., 2010. A method for evaluating dynamic functional network connectivity and task-modulation: application to schizophrenia. Magnetic Resonance Materials in Physics, Biology and Medicine 23, 351–366.

Scheinkman, J.A., LeBaron, B., 1989. Nonlinear dynamics and stock returns. The Journal of Business 62, 311–337. URL: http://www.jstor.org/stable/2353350.

Smith, S.M., Vidaurre, D., Beckmann, C.F., Glasser, M.F., Jenkinson, M., Miller, K.L., Nichols, T.E., Robinson, E.C., Salimi-Khorshidi, G., Woolrich, M.W., et al., 2013. Functional connectomics from resting-state fmri. Trends in cognitive sciences 17, 666–682.

Stoica, P., Moses, R.L., 2005. Spectral analysis of signals. Pearson/Prentice Hall Upper Saddle River, NJ.

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., Farmer, J.D., 1992. Testing for nonlinearity in time series: the method of surrogate data. Physica D: Nonlinear Phenomena 58, 77–94.

Van De Ville, D., Blu, T., Unser, M., 2004. Integrated wavelet processing and spatial statistical testing of fmri data. NeuroImage 23, 1472–1485.

Walker, G., 1931. On periodicity in series of related terms. Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 131, 518–532.

Wiener, N., 1930. Generalized harmonic analysis. Acta Mathematica 55, 117–258.

Yule, G.U., 1927. On a method of investigating periodicities in disturbed series, with special reference to wolfer's sunspot numbers. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 226, 267–298.

Zalesky, A., Fornito, A., Bullmore, E.T., 2010. Network-based statistic: identifying differences in brain networks. Neuroimage 53, 1197–1207.

Zalesky, A., Fornito, A., Cocchi, L., Gollo, L.L., Breakspear, M., 2014. Time-resolved resting-state brain networks. Proceedings of the National Academy of Sciences 111, 10341–10346.