Published in final edited form as:
Neuroimage. 2016 November 15; 142: 248–259. doi:10.1016/j.neuroimage.2016.05.023.

Untangling the Relatedness among Correlations, Part I: Nonparametric Approaches to Inter-Subject Correlation Analysis at the Group Level

Gang Chen[a,*], **Yong-Wook Shin**[b,†], **Paul A. Taylor**[a], **Daniel Glen**[a], **Richard C. Reynolds**[a], **Robert B. Israel**[c], and **Robert W. Cox**[a]

[a]Scientific and Statistical Computing Core, National Institute of Mental Health, National Institutes of Health, Department of Health and Human Services, USA

[b]University of Ulsan College of Medicine, Department of Psychiatry, Asan Medical Center, South Korea

[c]Mathematics Department, The University of British Columbia, Canada

Abstract

FMRI data acquisition under naturalistic and continuous stimuli (e.g., watching a video or listening to music) has become popular recently due to the fact that it entails less manipulation and more realistic/complex contexts involved in the task, compared to the conventional task-based experimental designs. The synchronization or response similarities among subjects are typically measured through inter-subject correlation (ISC) between any pair of subjects. At the group level, summarizing the collection of ISC values is complicated by their intercorrelations, which necessarily lead to the violation of independence assumed in typical parametric approaches such as Student's *t*-test. Nonparametric methods, such as bootstrapping and permutation testing, have previously been adopted for testing purposes by resampling the time series of each subject, but the quantitative validity of these specific approaches in terms of controllability of false positive rate (FPR) has never been explored before. Here we survey the methods of ISC group analysis that have been employed in the literature, and discuss the issues involved in those methods. We then propose less computationally intensive nonparametric methods that can be performed at the group level (for both one- and two-sample analyses), as compared to the popular method of circularly shifting the EPI time series at the individual level. As part of the new approaches, subject-wise (SW) resampling is adopted instead of element-wise (EW) resampling, so that exchangeability and independence assumptions are satisfied, and the patterned correlation structure among the ISC values can be more accurately captured. We examine the FPR controllability and power achievement of all the methods through simulations, as well as their performance when applied to a real experimental dataset. The new methodologies are shown to be both efficient and robust, and

*Corresponding author. gangchen@mail.nih.gov. †Corresponding author. shaman_korea@mac.com.

they have been implemented into an open source program, 3dNPT, in AFNI (http://afni.nimh.nih.gov).

Introduction

In a typical task-related FMRI experiment, the subject is presented with artificial stimuli or asked to perform meticulously-designed tasks in the scanner. Each experiment is categorized as an event-related design if each trial lasts for one TR or less, or as a block design if the duration of each trial is more than one TR; sometimes mixtures of the two are also performed. Through abstraction, simplification and reduction, the specification of various design features such as scanning parameters (TR, flip angle, etc.), explanatory variables (factors and quantitative covariates), and sample sizes (data points, repetitions of trials at the individual and number of subjects at the group level), allows the investigator to untangle various effects (main effects, interactions, various contrasts, trends, etc.) and to achieve proper statistical power in statistical inferences. In the analysis at the individual level, each voxel-wise time series is explained in a regression model through the construction of an idealized hemodynamic response (HDR) function or linear combinations of multiple basis functions, and the effect estimates are then taken to a general linear model at the group level to make generalizations about a population. However, in addition to (or, in part due to) the absence of distinctive textures of real life events, it has been argued (Hasson et al., 2004, 2008) that BOLD responses under such typical task FMRI paradigms (e.g., artificial or discrete intervals) are not as reliable as under naturalistically, continuously, and dynamically evolving conditions. Furthermore, the standard FMRI model-based approach through a presumed HDR function (Γ-variate in AFNI, Cohen, 1997; canonical function in SPM/FSL, Friston et al., 1998) may fail to capture the subtle shape differences in HDR across conditions or groups (e.g., Buxton et al., 2004; Barbé at al., 2012; Chen et al., 2015).

In light of these considerations, over the past decade it has been proposed (Hasson et al., 2004, 2008) that the presentation of a natural scene (e.g., a whole episode) during most of or the entire scanning session may be preferable. This “naturalistic” paradigm allows the investigator to explore the extent of synchronization, similarity, or shared processing at the same locations in the brain among subjects (e.g., Hasson et al., 2008), or even across species (Mantini et al., 2012). The differences in synchronization can be further explored across groups or conditions. Such a methodology with natural stimulus presentations has flourished and been applied to wide variety of FMRI experiments, such as visuoauditory movie stimuli (Hasson et al., 2004, 2008; Golland et al., 2007; Jääskeläinen et al., 2008; Kauppi et al., 2010; Nummenmaa et al., 2012), the synchronization of emotion (Nummenmaa et al., 2012), the impact of mass media coverage on various perceptions of the H1N1 pandemic (Schmälzle et al., 2013), real-world thought processing (e.g., educational television viewing of Sesame Street) between children and adults (Cantlon and Li, 2013), videos of dance performance (Herbec et al., 2015), narratives (Wilson et al., 2008), music (Abrams et al., 2013; Alluri et al., 2013; Thiede, 2014; Trost et al., 2015, Lillywhite et al., 2015), aesthetic performance (Jola et al., 2013), neural responses shared across languages (Honey et al., 2012), and political speeches (Schmälzle et al., 2015). In addition, applications have been seen in other neuroimaging modalities, such as MEG (Thiede, 2014), EEG (Bridwell et al., 2015), and ECoG (Potes et al., 2014). The approach has even been adopted by neuromarketers in the field of "neurocinematics" (Hasson et al., 2008).

The typical analysis approach to such naturalistic stimuli at the individual subject level involves computing the Pearson correlation coefficient of the EPI time series at each voxel of the brain in a standard space (e.g., Talairach-Tournoux or MNI) between any pair of subjects, leading to the terminology of inter-subject correlation (ISC) (Hasson et al., 2004). For one group with $n \geq 2$ subjects $S_1, S_2, \ldots, S_n$, there are $N=\frac{1}{2}n(n-1)$ correlation coefficients at each voxel from the ISC analysis, forming an $n \times n$ positive semi-definite (PSD) matrix,

$$\boldsymbol{R}^{(n)} = \begin{array}{c|c} & \begin{matrix} S_1 & S_2 & S_3 & \cdots & S_n \end{matrix} \\ \begin{matrix} S_1 \\ S_2 \\ S_3 \\ \vdots \\ S_n \end{matrix} & \begin{pmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1n} \\ r_{21} & 1 & r_{23} & \cdots & r_{2n} \\ r_{31} & r_{32} & 1 & \cdots & r_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & r_{n3} & \cdots & 1 \end{pmatrix} \end{array}, \tag{1}$$

where r_{ij} denotes the correlation between subjects i and j, $r_{ij} = r_{ji}$, $i \neq j$ (diagonals $r_{ii} = 1$, trivially). The symmetry of the ISC matrix means that we can focus on half of the off-diagonals, e.g., the N elements of the lower triangular part that is shaded in (1).

For statistical convenience, the correlation coefficients in (1) are usually converted to z-scores through the inverse hyperbolic tangent function *arctanh* (i.e., Fisher's Z-transformation), producing the symmetric matrix,

$$\boldsymbol{Z}^{(n)} = \begin{array}{c|c} & \begin{matrix} S_1 & S_2 & S_3 & \cdots & S_n \end{matrix} \\ \begin{matrix} S_1 \\ S_2 \\ S_3 \\ \vdots \\ S_n \end{matrix} & \begin{pmatrix} - & z_{12} & z_{13} & \cdots & z_{1n} \\ z_{21} & - & z_{23} & \cdots & z_{2n} \\ z_{31} & z_{32} & - & \ldots & z_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & z_{n3} & \cdots & - \end{pmatrix} \end{array}, \tag{2}$$

where z_{ij} represents the z-score between subjects i and j, $z_{ij} = z_{ji}$, $i \neq j$, and the dashes on the diagonal indicate that the transformation of 1 is indeterminate (and is of no interest). Similar to the conventional FMRI data analysis, the focus in the ISC analysis is to make proper generalization at the group level based on the N elements $\{r_{ij}, i > j\}$ or $\{z_{ij}, i > j\}$ in the lower triangular (shaded area in (1) or (2)) of the matrix $\boldsymbol{R}_n$ or $\boldsymbol{Z}_n$; that is, what is the typical ISC value and the associated statistical significance at the group level?

For two groups, G_1 and G_2, with subject sample sizes of n_1 and n_2 respectively in each group such that $n_1 + n_2 = n$, the ISC matrix can be partitioned as,

$$\boldsymbol{R}^{(n)} = \begin{array}{cc} & \begin{array}{cccccccc} \multicolumn{4}{c}{\overbrace{S_1 \qquad S_2 \qquad \cdots \qquad S_{n_1}}^{G_1}} & \multicolumn{4}{c}{\overbrace{S_{n_1+1} \qquad S_{n_1+2} \qquad \cdots \qquad S_n}^{G_2}} \end{array} \\ \begin{array}{r} G_1\left\{\begin{array}{c} S_1 \\ S_2 \\ \vdots \\ S_{n_1} \end{array}\right. \\ G_2\left\{\begin{array}{c} S_{n_1+1} \\ S_{n_1+2} \\ \vdots \\ S_n \end{array}\right. \end{array} & \left(\begin{array}{cccc|cccc} 1 & r_{1,2} & \cdots & r_{1,n_1} & r_{1,n_1+1} & r_{1,n_1+2} & \cdots & r_{1,n} \\ r_{2,1} & 1 & \cdots & r_{2,n_1} & r_{2,n_1+1} & r_{2,n_1+2} & \cdots & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n_1,1} & r_{n_1,2} & \cdots & 1 & r_{n_1,n_1+1} & r_{n_2,n_1+2} & \cdots & r_{n_1,n} \\ \hline r_{n_1+1,1} & r_{n_1+1,2} & \cdots & r_{n_1+1,n_1} & 1 & r_{n_1+1,n_1+2} & \cdots & r_{n_1+1,n} \\ r_{n_1+2,1} & r_{n+2,2} & \cdots & r_{n_1+2,n_1} & r_{n_1+1,n_1+2} & 1 & \cdots & r_{n_1+2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & r_{n,n_1} & r_{n,n_1+1} & r_{n,n_1+2} & \cdots & 1 \end{array}\right), \end{array} \tag{3}$$

where the three colors, blue, red, and green, correspond, respectively, to the three partitions, $\boldsymbol{R}_{11}$, $\boldsymbol{R}_{22}$, and $\boldsymbol{R}_{21}$, of the N elements in the lower triangular part of $\boldsymbol{R}^{(n)}$ (see schematic representations in Fig. 1): two within-group correlation (WGC) subsets $\boldsymbol{R}_{11}$ and $\boldsymbol{R}_{22}$ for groups G_1 and G_2, respectively, as well as the between-group correlation (BGC) subset $\boldsymbol{R}_{21}$. The corresponding z-matrix $\boldsymbol{Z}_n$ can be partitioned in the same fashion.

Here we focus on separate prototypes of statistical tests for ISC data with one and two groups of subjects, corresponding to the conventional one- and two-sample t-tests, respectively. We note that the paired t-test can be reduced to a one-sample t-test, and comparisons among multiple groups can be performed using multiple two-sample t-tests. Using the framework introduced here, there are two types of between-group analysis – a standard "direct" comparison between the two WGC subsets $\boldsymbol{R}_{11}$ and $\boldsymbol{R}_{22}$, and a novel "indirect" comparison for the difference between a WGC component $\boldsymbol{R}_{11}$ (or $\boldsymbol{R}_{22}$) and the BGC subset $\boldsymbol{R}_{21}$ (Fig. 1).

One major concern affecting the choice of statistical approach with ISC data at the group level is that $\{r_{ij}, i > j\}$ (or, equivalently, $\{z_{ij}, i > j\}$) are not statistically independent samples. Previously, both parametric and nonparametric methods have been seen in the literature and, for example, implemented into an analytical package ISC toolbox in Matlab (Kauppi et al., 2014; https://www.nitrc.org/projects/isc-toolbox/). Here, we first strive to explicitly characterize the relatedness among the ISC values through a structured variance-covariance matrix described in the Methods section, and we then propose a set of new nonparametric methods in light of the variance-covariance structure. We further compare these, along with various existing approaches, by analyzing ISC data in their ability to control false positive rates (FPRs) using simulations, and we explore the performances of each approach on a real experimental dataset. This paper mainly focuses on nonparametric approaches to ISC group analysis, while an upcoming article, referred to as "Part II" hereafter, will explore the applicability of a novel parametric method in handling the ISC variance-covariance structure; the complementary parametric methods will be shown to have the additional capability for estimating several of the parameters introduced here, as well as providing further, useful interpretations for them.

Methods: Theory

ISC variance-covariance structure

Throughout this article, regular italic letters in lower (e.g., a) and upper (Z) case stand for scalars and random variables, respectively; boldfaced italic letters in lower ($\boldsymbol{a}$) and upper ($\boldsymbol{X}$) cases for column vectors and matrices, respectively. Major acronyms used in the paper are tabulated in Appendix A. Suppose that z_{ij} and z_{kl} are two z-values associated with the ISCs of r_{ij} and r_{kl}, and Z_{ij} and Z_{kl} are the corresponding random variables[1] from which z_{ij} and z_{kl} are sampled or instantiated. Let ψ denote the correlation between Z_{ij} and Z_{kl} that pivot around one subject (i.e., the two index pairs (ij, kl) share one and only one common index such as (3 2, 4 2) or (3 1, 4 3)). In other words, ψ characterizes the interrelatedness of Z_{ij} and Z_{kl} among three subjects. It is reasonable to assume that the correlation matrix $\boldsymbol{P}^{(n)}$ for $\{Z_{ij}, i > j\}$ has the following structure,

$$corr(Z_{ij}, Z_{kl}) = \begin{cases} 1, & \text{if the cardinality of } \{i, j\} \cup \{k, l\} \text{ is } 2 \text{ (i. e. , } i=k, j=l); \\ \rho, & \text{if the cardinality of } \{i, j\} \cup \{k, l\} \text{ is } 3 \text{ (i. e. , there are only three distinct indices)}; \\ 0, & \text{if the cardinality of } \{i, j\} \cup \{k, l\} \text{ is } 4 \text{ (i. e. , the four indices are distinct)}. \end{cases} \quad (4)$$

We further define $\boldsymbol{z} = vec(\{Z_{ij}, i > j\})$ to be the vector of length N whose elements are the column-stacking of the lower triangular part of the matrix $\boldsymbol{Z}^{(n)}$ in (2) or its two-sample version. That is, $\boldsymbol{z}$ is the half-vectorization of $\boldsymbol{Z}^{(n)}$ excluding the main (or principal) diagonal: $\boldsymbol{z} = vech(\boldsymbol{Z}^{(n)})\backslash diag(\boldsymbol{Z}^{(n)})$. The variance-covariance matrix of $\boldsymbol{z}$ can be expressed as[2] the $N \times N$ matrix,

$$\sum\nolimits^{(n)} = \sigma^2 \boldsymbol{P}^{(n)}, \quad (5)$$

[1]The random variable Z_{ij} can be conceptually thought of as representing the possible values that z_{ij} may take. Mathematically we may define it as a measurable function that maps two EPI time series to the real-valued space $\mathbb{R}$.

[2]If the two voxel-wise time series follow a bivariate Gaussian distribution, Z_{ij} can be approximated by a Gaussian distribution $G(arctanh(r), \sigma^2)$, $\sigma^2 = \frac{1}{T-k-3}$, where r is the true correlation coefficient, T is the number of data points in the voxel-wise time series, and k is the number of confounding effects accounted for in the ISC analysis.

where σ^2 is the variance of Z_{ij}, $i > j$. For example, with $n = 5$ subjects, the correlation matrix $\boldsymbol{P}^{(n)}$ for $\boldsymbol{z}$ with the $N=\frac{1}{2}n(n-1)=10$ ISC z-values has the following form,

$$\boldsymbol{P}^{(5)}= \begin{array}{c} \\ Z_{21} \\ Z_{31} \\ Z_{41} \\ Z_{51} \\ Z_{32} \\ Z_{42} \\ Z_{52} \\ Z_{43} \\ Z_{53} \\ Z_{54} \end{array} \begin{array}{cccccccccc} Z_{21} & Z_{31} & Z_{41} & Z_{51} & Z_{32} & Z_{42} & Z_{52} & Z_{43} & Z_{53} & Z_{54} \\ \hline \end{array} \hspace{-0pt} \begin{pmatrix} 1 & \rho & \rho & \rho & \rho & \rho & \rho & 0 & 0 & 0 \\ \rho & 1 & \rho & \rho & \rho & 0 & 0 & \rho & \rho & 0 \\ \rho & \rho & 1 & \rho & 0 & \rho & 0 & \rho & 0 & \rho \\ \rho & \rho & \rho & 1 & 0 & 0 & \rho & 0 & \rho & \rho \\ \rho & \rho & 0 & 0 & 1 & \rho & \rho & \rho & \rho & 0 \\ \rho & 0 & \rho & 0 & \rho & 1 & \rho & \rho & 0 & \rho \\ \rho & 0 & 0 & \rho & \rho & \rho & 1 & 0 & \rho & \rho \\ 0 & \rho & \rho & 0 & \rho & \rho & 0 & 1 & \rho & \rho \\ 0 & \rho & 0 & \rho & \rho & 0 & \rho & \rho & 1 & \rho \\ 0 & 0 & \rho & \rho & 0 & \rho & \rho & \rho & \rho & 1 \end{pmatrix}$$

Among the N lower triangular elements in $\boldsymbol{Z}^{(n)}$, $n - 2$ of them carry index i (or j), excluding the term z_{ij} for each ij index pair ($i > j$). Therefore, the numbers of ψ and 0 in each row (or column) in the matrix $\boldsymbol{P}^{(n)}$ are $2(n - 2)$ and $\frac{1}{2}(n-2)(n-3)$, respectively. While the correlation amplitude ψ is generally unknown a priori, the necessary and sufficient condition to guarantee the PSD property of $\boldsymbol{P}^{(n)}$ with $n > 3$ is (Appendix B)

$$-\frac{1}{2(n-2)} \leq \rho \leq 0.5. \qquad (6)$$

Despite including the small, extra interval of negative values $[-\frac{1}{2(n-2)}, 0)$, which uniformly vanishes as n increases, the domain of interest for ψ here is [0, 0.5], because negative correlation between two Z_{ij} samples from each subject, while mathematically possible, is highly unlikely in reality. Moreover, the two boundary values are associated with two extreme scenarios: $\psi = 0$ corresponds to the situation of complete independence among the z_{ij}, while $\psi = 0.5$ means that any two z-values of a subject are maximally correlated, which would correspond to the special case of no noise in a parametric model (see Part II).

The independence assumption in the conventional parametric method such as Student's t-test would be violated for dataset $\{z_{ij}, i > j\}$, with the nontrivial matrix of $\boldsymbol{P}^{(n)}$ (i.e., it is not necessarily the identity matrix) characterizing the degree to which the assumption of independence is violated unless $\psi = 0$. In other words, even with the assumption of multivariate Gaussianity for $\boldsymbol{z}$, it would still be a challenge to appropriately handle the variance-covariance matrix $\boldsymbol{\Sigma}^{(n)} = \sigma^2\boldsymbol{P}^{(n)}$ under conventional parametric paradigms. On the other hand, nonparametric methods such as permutation testing and bootstrapping become desirable because they can enable inferences on the statistic of interest when: a) the true distribution of this statistic (e.g., the structure in $\boldsymbol{P}^{(n)}$) is unknown or too complicated, or b) only a weaker assumption than independence is required. The crucial mechanism by which the nonparametric method acts is the generation of a null distribution against which the observed effect of interest can be inferred. Notice that, since nonparametric methods do not assume a particular distribution, it is not necessary to use the Fisher-transformed ISC, and we can focus on the original correlation coefficients $\{r_{ij}, i > j\}$ instead of $\{z_{ij}, i > j\}$. Additionally, we can adopt the median as the chosen centrality measure for ISC, instead of the mean as in parametric methods, which has mathematical and practical advantages (see Discussion).

Permutations

In general, permutation testing only requires that exchangeability is satisfied when randomly permuting (or reassigning) the units; that is, exchangeability assumes independence between the potential outcomes and the permutation (or assignment) mechanism (i.e. each unit is not differentiable from other units before the units are assigned). In the case of having only one group of subjects, the typical permutation strategy is to randomly flip the sign for some members; for two groups, one can establish the exchangeability by randomly shuffling the membership between the groups. In the end, the magnitude of the chosen centrality measure of the observed data (e.g., median) can be statistically inferred against the null distribution generated through an appropriate number of permutations (e.g., 5000).

The crucial choice to be made is how to select the pivoting units. If we perform element-wise (EW) permutation (EWP) by pivoting around the N individual elements of $\{r_{ij}, i > j\}$, the exchangeability would be broken because: a) the integrity of the correlation matrix $\boldsymbol{R}^{(n)}$ (e.g, its PSD property) may be lost; and b) each element in $\{r_{ij}, i > j\}$ or each slot in the shaded area of $\boldsymbol{R}^{(n)}$ in (1) is not non-differentiable when permuting, as characterized by the pattern in the correlation structure $\boldsymbol{P}^{(n)}$ when $\psi \neq 0$. On the other hand, if we execute subject-wise (SW) permutation (SWP) by pivoting at the subject level (e.g., flipping signs in the case of one group, or shuffling group memberships - exchanging rows and columns in $\boldsymbol{R}^{(n)}$ - between two groups), we will be able to maintain the integrity of the resulting correlation matrix as well as the independence between the permuted correlation matrices. In other words, general exchangeability can only be met with SWP, and not with EWP, and the differences are exemplified in the EWP and SWP rows of Table 1.

Two additional features of SWP are worth noting here. First, in the case with one group, double sign flipping, which is equivalent to no sign change, happens under SWP where a row and column of two flipped subjects intersect (e.g., for elements r_{41}, r_{61}, r_{64} on the row SWP of the "One Group" column in Table 1 where subjects S_1, S_4, and S_6 have been flipped). Second, a unique feature of SWP with two groups is that the BGC values (from the green submatrix in (1), $\boldsymbol{R}_{21}$ in Fig. 1) are involved in generating the null distribution, as illustrated[3] with the green correlation values in the "SWP, Two Groups" case in Table 1.

Bootstrapping

Bootstrapping is an alternative nonparametric approach that resamples the data with replacement within each group so that each resampled dataset has the same size as the original one. It employs the assumption that the observed data are randomly drawn from an

[3]Even though it is demonstrated in Table 1 with equal number of subjects across the two groups, such a constraint of balance is not required.

independent and identically distributed population, from which it estimates the distribution of the observed data. Applied to the ISC context, we can again choose to focus on the centrality of each resampled dataset (e.g., median) at each iteration, and, with enough number of repetitions (e.g., 5000), obtain an empirical population distribution from which a null distribution is constructed through shifting the sampled distribution by the observed centrality measure. In the same vein as in permutation testing, we may select from two choices of resampling: element-wise bootstrapping (EWB) and subject-wise bootstrapping (SWB). In the manner as described above in the permutation case, so does EWB violate the resampling assumption of independence, while SWB does not.

When applied to ISC data, bootstrapping with replacement for SWB may lead to a situation with one subject occurring more than once, leading to an off-diagonal correlation coefficient of 1 (e.g., highlighted with squares in the "SWB" row in Table 1). The appearance of this feature necessitates the direct use of correlation coefficients $\{r_{ij}, i > j\}$ instead of their transformed z-values, as the transformation value is undefined when $r = 1$. Both simulations and real experimental data are used in the next section to show the consequences of the various assumptions and the measurable differences of each approach.

Simulations and Real Experiment Results

Simulations of group analysis with different testing methods

We performed simulations in a 2 × 4 × 6 × 5 factorial design with our focus on:

- **a.** 2 types of ISC group analysis: one- (one group) and two-sample (two groups);
- **b.** 4 sample sizes: 10, 20, 40, and 80 subjects in each group;
- **c.** 6 parameter values: Six ψ values were selected from the interval of [0, 0.5] with a step size of 0.1; and
- **d.** 5 testing methods: 1) Student's t-test (T), 2) element-wise permutations (EWP), 3) element-wise bootstrapping (EWP), 4) subject-wise permutations (SWP), and 5) subject-wise bootstrapping (SWB). For the two-sample scenario, three comparisons were performed: direct contrast between the two WGCs ($\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{22}$), and indirect contrast of each WGC versus the BGC ($\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{21}$ and $\boldsymbol{R}_{22}$ vs $\boldsymbol{R}_{21}$).

To examine the FPR controllability and power attainment for each of the 2×4×5 scenarios, 5000 simulated datasets were generated, each of which was an instantiation of $\boldsymbol{z}$, containing N values[4] of $\{z_{ij}\}$ drawn from an N-variate Gaussian distribution $G(\boldsymbol{\mu}, \Sigma)$, with the variance-covariance matrix Σ defined per the structure (5), where $\sigma^2 = 1$. For FPR, $\boldsymbol{\mu}_N = \boldsymbol{0}_N$ for all cases, while for power analyses, $\boldsymbol{\mu}_N = 0.5 \cdot \boldsymbol{1}_N^T$ (one group) or $\boldsymbol{\mu}_N = (\boldsymbol{0}_{N-N_2}^T, 0.5 \cdot \boldsymbol{1}_{N_2}^T)^T$ (two groups, $N_2 = \frac{1}{2} n_2(n_2 - 1)$ is the number of elements in $\boldsymbol{Z}^{(n)}$ for the second group). The z-

[4]For these simulations, data $\{z_{ij}\}$ (the instantiations of Z_{ij}) were drawn from a Gaussian distribution, and r_{ij} were calculated using the inverse Fisher transformation; in practice, when analyzing ISC data non-parametrically, one would calculate the correlation coefficients r_{ij} directly from the EPI time series.

values from the 5000 simulated datasets were directly adopted for the parametric test (i.e., Student's *t*), but were inverse Fisher-transformed (to mimic actual correlation ceofficients) through hyperbolic tangent function *tanh* before being fed into the nonparametric methods, in which the median was the selected centrality measure. The FPR for each scenario was estimated by counting the number of realizations out of the simulated 5000 datasets that reached the nominal significance level of 0.05.

The FPR and power estimates for each method are shown in Fig. 2. For the one-sample case (upper row in Fig. 2), the FPR increases across all the five methods when the correlation parameter ψ goes from 0 to 0.5 across all the sample sizes. When $\psi = 0$, all the methods except for SWB are aligned with the nominal FPR of 0.05. This phenomenon is expected, especially for the Student's *t* (red), EWB (orange), and EWP (black diamonds) because the ISC values of a subject with two other subjects are independent of each other when $\psi = 0$. On the other hand, for $\psi = 0$ SWB (blue) is too conservative: in this special case, the "bundling" of ISC values per subject, when they are actually independent of each other, leads to overly conservative identifications. As ψ increases from 0, all the methods except for SWB quickly become too liberal. In contrast, even though moderately liberal when ψ is close to 0.5, SWB performs reasonably well for a broad range of ψ, particularly as the number of subjects increases.

For the two-sample case (second row in Fig. 2), permutation testing at the subject level (SWP, black) is uniformly well-behaved in terms of FPR controllability across all sample sizes and across the whole range of ψ values. In contrast, permutations at the element level (EWP, purple) tended to be too conservative when the correlation parameter ψ is low and too liberal when ψ is above 0.2. All the other three methods (Student's *t*, EWB, and SWB) show expected over-conservative controllability of FPR when $\psi = 0$, because this represents the special case in which all the ISC values are independent of each other, even from the same subjects. While EWB and Student's *t* demonstrate unacceptable FPR controllability, SWB for two groups shows similar behavior to the one group case, and offers an alternative to SWP even though it fares less well with a small sample size or when $\psi \leq 0.1$. Unlike the case with one group, in which having the permutations at the subject level does not appear to provide much gain relative to its EW counterpart, the BGC values (the elements in the $\boldsymbol{R}_{21}$ submatrix) in the two-group case provide a crucial leverage for the null distribution, leading to the well-balanced performance of SWP.

We emphasize the dramatically different performance between SWP for two-group comparison and all other cases, including SWP with one group: SWP controls FPR uniformly well across all ψ values and sample sizes for the two-group comparison, while FPR for all other cases monotonically increases as the extent of correlation among $\{z_{ij}\}$ increases. The essentially ideal FPR performance and the immunity of SWP to the magnitude of ψ in the two-group case are achieved probably through the combination of two unique features: the subject-wise execution and the involvement of BGC (green elements in (3), $\boldsymbol{R}_{21}$ of Fig. 1), the latter of which is absent from all other scenarios.

The following overall trends are evident across all the methods in power performance. For a given ψ, power increases monotonically with sample size in all cases, as expected. When ψ

= 0, the power of each method reaches the highest because full independence allows for relatively easy detection among the elements. In contrast, when $\psi = 0.5$, the power is the lowest because it is most difficult to tease apart the intercorrelations among the elements and to eliminate false positives. Consistent with the FPR controllability, SWB produced the lowest power with one group, while the other four methods had roughly equivalent power attainment. Similarly for two groups, the order of power achievement was reversed relative to their FPR performance. Also similar to their FPR comparison, SWP outperformed/ underperformed relative to EWP when ψ was low/high.

In summary, all the methods, except SWP with two groups, render worsening FPR control as the independence assumption becomes more severely violated albeit to varying degrees. For comparing two groups SWP is ideal and SWB can be a suboptimal alternative, while all other methods do not fare well in FPR. With one group, SWB offers the best solution among the alternatives for one group: it controls FPR reasonably well when the amount of correlation is moderate, but it does tend to be slightly conservative/liberal when ψ is small/ big. On the other hand, the power monotonically increases for all methods with a bigger sample size, and all the methods had reversed order in the power relative to their FPR performance; the power gain comes at the cost of poor FPR control.

Performance comparisons with experimental data

To demonstrate and compare the various modeling approaches for ISC analysis at the group level in real FMRI data, we utilize here the same experimental data used in Shin et al. (under review). Briefly, $n = 48$ healthy volunteers ($n_1 = 24$ males, $n_2 = 24$ females, age mean ± SD = 33.6 ± 5.7 and 34.7 ± 6.0 years old for males and females, respectively) watched six movie clips, each with an average length of two and half minutes, in a 3.0-T Siemens Trio scanner. Half of the six clips depicted mostly positive emotional episodes while the other half were of negative emotional valence. The series of clips were separated by a black screen for 10–30s and preceded by a fixation cross for 30s, leading to a total scanning time of 1,050 seconds. Parameters for the acquired whole brain BOLD EPI data were: voxel size of 3.8 × 3.8 × 4.0 mm^3, 36 axial slices, TR = 2,000 ms, TE = 30 ms, in plane FOV = 240 × 240 mm^2, flip angle = 90°.

The EPI time series went through the following preprocessing steps in AFNI (Cox, 1996) using afni_proc.py: de-spiking, slice timing and head motion corrections, affine spatial alignment to a Talairach template (TT_N27) at a voxel size of 2.0×2.0×2.0 mm^3, smoothing with an isotropic FWHM of 8 mm, and removal of physiological noise such as cardiac and breathing effects using ANATICOR (Jo et al., 2010). ISC was computed over 205 time points (having excluded the periods of fixation and blank screen) at the voxel level between all pairs of the $n = 48$ subjects using 3dTcorrelate in AFNI, leading to $N = 48 \times 47/2 = 21,128$ ISC values per voxel.

Of interest here were statistical inferences of the ISC for each group (i.e., one-sample tests for each of $\boldsymbol{R}_{11}$ and $\boldsymbol{R}_{22}$), the difference between the two sexes (direct comparison of WGC, $\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{22}$), and each WGC in contrast to the BGC (indirect comparisons of $\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{21}$ and $\boldsymbol{R}_{22}$ vs $\boldsymbol{R}_{21}$); that is, five tests: two within-group, one between-group comparison, and two within-versus between-group, were performed using AFNI programs 3dttest++ (for

Student's *t*) and 3dNPT (for EWP, SWP, EWB, and SWB). The results of the within-group (male) and between-group comparison are illustrated in Fig. 3 while the SWP performance for the other two comparisons ($\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{21}$ and $\boldsymbol{R}_{22}$ vs $\boldsymbol{R}_{21}$) are shown in Fig. 4 of Appendix C. The computation time for the nonparametric methods was approximately 6 hours for each group analysis case of one group ($n = 24$ subjects with 12 CPUs) and two groups ($n = 48$ subjects with 24 CPUs) on a Linux system (Fedora 14) with Intel® Xeon® X5650 at 2.67GHz.

We note the following two aspects of the results with our real experimental data. First, the performances of the five methods were consistent with the FPR controllability and power achievement from the simulations in Fig. 2. As typically seen in the ISC literature, the volume of statistical significance for a single group was overwhelming even for a small *p*-value threshold ($p \leq 0.001$, upper panel, Fig. 3A). In contrast, the two-group comparison (lower panel, Fig. 3A) was similar to the conventional FMRI results in the sense that the statistically significant regions are largely localized. Relative to SWB, all the other five methods are largely inflated for one group (upper panel in Fig. 3A and voxel counts in Fig. 3B). For two-group comparison, Student's *t*-test and EWB had the worst inflation (lower panel in Fig. 3A and voxel counts in Fig. 3B).

Secondly, we note that the full set of results in Fig. 3 indicate that $\psi > 0$, otherwise all the methods except SWB (or EWP) would have roughly the same FPR and detection power, corresponding to similar thresholded volumes, for the case of single group or two-group comparison with 20 or more subjects (*cf.* Fig. 2). In other words, the special case of having independence among the ISC values does not appear to generally hold in real data, and therefore it must be accounted for in both modeling and analysis. More subtly, as shown in the simulations above, SWP was more/less powerful than EWP probably depending on the smaller/larger magnitude of the correlation parameter ψ, which is consistent with the simulation results of SWP vs. EWP (*cf.* the intersecting lines of black and purple on the second and fourth rows in Fig. 2). Furthermore, SWP is largely more powerful than SWB in comparing two groups (voxel counts in Fig. 3B), suggesting $0 < \psi < 0.2$ (*cf.* the intersecting lines of black and blue in the second column on the second and fourth rows in Fig. 2) in most - but not all - regions in the brain. This deduction will be further confirmed by the estimations of ψ through a parametric approach in Part II.

For direct comparisons, we applied the permutation approach implemented in the ISC Toolbox (version 2.1, using the recommended default with 100 million randomizations; Kauppi et al., 2014) in Matlab (version R2015b) to the male group of 24 subjects (two group comparison is currently not available in the toolbox). The runtime with 64 CPUs and 1 TB of RAM was 34.7 hours on a 64-bit CentOS server (Intel® Xeon® X7560 at 2.27GHz). The result in the sixth column in Fig. 3A shows the output from the ISC Toolbox, which is the average voxel-wise ISC *r*-value (unlike the results from other nonparametric tests, which are median *r*-values; the output from the *t*-test is the inverse transform of the mean *z*-values). The large size of the suprathreshold volume (Fig. 3A) and the number of suprathreshold voxels (Fig. 3B) suggest that the approach of permuting at the time series level in the ISC Toolbox led to similarly inflated significance and poor FPR control as EWP, SWP, and EWB.

Discussion

Compared to the conventional task-related FMRI experiments in which the investigator usually assumes a pre-fixed HDR function, the paradigm with naturalistic stimuli usually elicits more robust results from a model-free ISC analysis without involving any prior knowledge about HDR (Hansson et al., 2008a). To make generalizations about ISC-based inference at the group level, a major challenge is the fact that, even though the n subjects involved are independent samples, their $N = n(n-1)/2$ ISC values $\{r_{ij}, i > j\}$ are typically not independent. This interdependence is characterized here by the correlation structure of $\boldsymbol{P}^{(n)}$ with ψ having a value within $[-\frac{1}{2(n-2)}, 0.5]$, where independence of the N ISC values would be represented by the special circumstance of $\psi = 0$. The comparisons of the methods in the real experimental data analyzed here (Fig. 3) strongly suggest that the value of ψ is both positive and non-negligible, and so the interrelatedness of ISC values needs to be accounted for in any group analysis. We emphasize that the correlation structure $\boldsymbol{P}^{(n)}$, as an assumption and parameterization for $\{z_{ij}, i > j\}$, is not required for actually implementing thenonparametric methods, but it has been motivated and presumed here in order to provide a framework for discussion and for our simulations (Fig. 2). We note that ψ was introduced here as being constant across a group, which reasonably reflects the basic fact that, in general, subjects are recruited as statistically equivalent representatives of a population; moreover, this is consistent within the context of nonparametric testing, as all units in the permutation tests are assumed to be indistinguishable and exchangeable. Additionally, comparisons of the simulation and real data result strongly suggest that ψ is nonzero in practice, meaning that there is nontrivial variance-covariance structure across a typical group in FMRI studies. These assumptions and structures will be further explored with parametric approaches in Part II, which offers further means to interpret and even estimate these structural parameters.

The typical FMRI parametric group analysis tools such as Student's t-test, ANOVA or GLM would have difficulty in accurately handling the special correlation structure of $\boldsymbol{P}^{(n)}$, which has been shown to be highly likely to exist within a group in practice. This difficulty is evidenced by the huge inflations of FPR in the simulations and real experimental results presented here. Instead, nonparametric methods such as permutation testing have long been adopted in neuroimaging (Nichols and Holmes, 2002; Mériaux et al., 2006; Winkler et al., 2014; Winkler et al., 2015). In the context of ISC data, they become a natural choice thanks to their meager assumptions about data distribution. However, the intercorrelations among $\{r_{ij}, i > j\}$ or $\{z_{ij}, i > j\}$ remain a hurdle even for nonparametric approaches. For example, exchangeability is required when resampling the data with permutations. For bootstrapping, the observations - ISC values in our case - are assumed to come from an independent and identically distributed population, in addition to the assumption of independent resampling. We have presented and evaluated new approaches that meet these requirements for the nonparametric methods.

In addition to the advantages of having less human intervention in the tasks, natural and complex stimuli may reveal more reliable, selective and time-locked responses that might be difficult to detect in the conventional FMRI experiments (Hasson et al., 2010). Such an experimental design could be generally applied more broadly to paradigms that contain blocks of data under one task or condition. The consideration of the correlation structure represented by $\boldsymbol{P}^{(n)}$, and the subject-wise permutations/bootstrapping, would be equally relevant in these situations, as well.

Methodology survey

In the early days, the pioneering work with naturalistic stimuli were performed either within each subject when the natural stimulus was repeated several times (Hasson et al., 2008b) or through ISC for each subject pair separately without summarization at the group level (Hasson et al., 2004), in which case the ISC results were typically verified through seed-based correlation analysis (Hasson et al., 2004; Hasson et al., 2008b; Schmälzle et al., 2013). Later on, some investigators simply ran one-sample (Bartels and Zeki, 2004; Hasson et al., 2008a; Wilson et al., 2008; Abrams et al., 2013; Kauppi et al., 2014), two-sample (Schmälzle et al., 2013; Cantolon and Li, 2013) or paired (Abrams et al., 2013; Schmälzle et al., 2015) t-tests on z-values $\{z_{ij}, i < j\}$ of correlation coefficients, while it was generally acknowledged that the N elements $\{z_{ij}, i > j\}$ were not independent, as illustrated in the correlation structure of $\boldsymbol{P}^{(n)}$ in (4), thereby violating the independence assumption in the Student's t-test and leading to the inflated degrees of freedom for the t-distribution. The approach was mainly justified based on the observation that the null results generated by shifting each pair of time series by random steps roughly fitted to a $t(N-1)$-distribution curve (Wilson et al., 2008).

Nonparametric methods have also been adopted in the previous ISC literature. For example, one popular approach for one condition was to acquire a null distribution for the whole brain by randomizing the time series across voxels and time points (e.g., circularly shifting each subject's time series by a random lag so that they were no longer aligned in time across the subjects) (Kauppi et al., 2010; Nummenmaa et al., 2012; Abrams et al., 2013; Kauppi et al., 2014; Pajula and Tohka, 2014; Trost et al., 2015; Herbec et al., 2015; Bridwell et al., 2015). For comparing two conditions, a different permutation test was used on the N elements $\{z_{ij}, i < j\}$ through element-wise sign flipping (Kauppi et al., 2014). Yet another permutation strategy was to generate a null distribution of group-wise ISCs by randomly shuffling the subjects between two groups at the ROI level (Schmälzle et al., 2013). One variation of this ISC approach is to first calculate the ISC value of a subject between a voxel's BOLD time course in the subject and the average of that voxel's BOLD time course in the remaining subjects (Kauppi et al., 2010; Honey et al., 2012; Schmälzle et al., 2013; Schmälzle et al., 2015). Then, at the group level, EWB was adopted to make inferences through phase-randomization in the EPI time series (Honey et al., 2012; Schmälzle et al., 2015). Bootstrapping was also implemented through scrambling the frequency bands and channels in ECoG data (Potes et al., 2014). Some of these methods discussed above have been implemented in a publicly available analytical package ISC Toolbox (Kauppi et al., 2014); however, the results produced by this toolbox with the real experimental data here indicates that its FPR is likely to be largely inflated (Fig. 3), just as EWB, EWP, and SWP are.

Relatedly, permutation testing has been utilized in the context of a distance matrix, where the distance of two random variables with a correlation coefficient r is defined as, for example, $\sqrt{2(1-r)}$ (Gower and Krzanowski, 1999). The distance matrix is then analyzed through multivariate distance matrix regression (MDMR) with the flexibility of incorporating multiple between-subject explanatory variables such as groups and quantitative covariates (Anderson, 2001; Zapala and Schork, 2006; Reiss et al., 2010). More relevantly, MDMR has been applied to voxel-wise distance matrices of two groups that were computed from the seed-based correlation of resting-state data in gray matter (Shehzad et al., 2014). It would be intriguing to see how MDMR, when applied to ISC data, compares to the nonparametric methods explored here in terms of FPR and power achievement.

Issues with previous ISC group analysis methodology

Group ISC average—One common practice in the literature is to obtain the group mean directly by averaging the ISC values across N subject pairs $\{r_{ij}, i > j\}$ (Kauppi et al., 2010; Nummenmaa et al., 2012; Honey et al., 2012; Pajula and Tohka, 2014; Kauppi et al., 2014; Bridwell et al., 2015; Lillywhite et al., 2015; Herbec et al., 2015). However, mathematically, the averaging as implemented in the ISC Toolbox (Kauppi et al., 2014) is problematic for the following reason. Correlation coefficients are not additive in the sense that correlation is not a linear function of the strength of the relation between the two random variables (or time series, in the context of FMRI data). Simply averaging them would lead to biases or underestimation in representing the whole group in the sense that estimates tend to be closer to 0 whether positive or negative (Silver and Dunlap, 1987). Specifically, averaging among correlation coefficients renders the interpretation of the final result difficult or even without technical meaning. As the correlation coefficient between two random variables X and Y is defined as $r=\frac{Cov(X,Y)}{\sqrt{Var(X)\,Var(Y)}}$ X and Y or $r=\frac{1}{n-1}\sum\frac{(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y}$ between two time series $\{x_i\}$ and $\{y_i\}$, averaging the r-values across subjects is equivalent to averaging the covariance and the two variances across subjects. This cannot be justified because the group variance and covariance are usually larger than or different from the weighted sum of variances and covariances (due to between-subject differences) unless, very specifically, the means of all subjects on both variables or time series are exactly the same.

The bias introduced through simple averaging can be demonstrated using an arithmetic example: With two correlation coefficients $r_1 = 0.5$ and $r_2 = 0.9$, the different results from the simple average and from Fisher's transformation lead to estimates of 0.7 and 0.77, respectively, showing a greatly skewed value for the simple averaging toward r_1. Such biased estimates were also revealed in the output from the ISC Toolbox (Fig. 3, though hardly identifiable from the color coding).

Instead of simple averaging, what is needed for summarizing is to approximately Gaussianize the correlation coefficients through transformation (e.g., Fisher's Z or squaring to the coefficients of determination, R^2). A more rigorous approach would be to go through three steps: Fisher's Z-transformation, averaging, and converting the z-values back to correlation coefficient. Yet another alternative, as implemented here with our bootstrapping and permutation testing, is to use a centrality measure such as median, eschewing the back-and-forth transformation processes.

Difficulty of validations—As noted above, the basic assumption of statistical independence in several of the existing methods surveyed above is largely questionable when applying Student's *t*-tests. Additionally, a strict assessment of some methods such as permutations or bootstrapping at the level of EPI time series remains elusive because it is difficult to assess their ability in controlling FPR. That is, whether the current approaches properly account for the correlation structure represented in $\boldsymbol{P}^{(n)}$ remains unaddressed. As shown here, methods which ignore the underlying correlation structure tend to be highly susceptible to having an inflated FPR. It is difficult, if not unlikely, to examine FPR controllability with simulated time series data at the individual subject level. For example, time series with simple white noise or even autoregressive or autoregressive-moving-average models would correspond to $\psi = 0$ in (4), leading to unrealistic simulations. Another potential issue for some permutation tests is the procedure of circularly shifting the sequence of the EPI time series at the individual level: it remains unclear whether the exchangeability criterion is met, considering the fact that the temporal integrity is kept through the process.

Without direct evaluation of FPR controllability, the validity (e.g., inflated significance) of the results using any method remain undetermined. In the ISC literature, one approach to dealing with the resultant widespread detections (Hasson et al., 2004; Schmälzle et al., 2013; see also the upper panel in Fig. 3), which are potentially overinflations due to poor FPR control, has been to apply a secondary conservative safeguard against false positives. For example, Bonferroni correction or an extremely low false discovery rate (e.g., $q = 10^{-7}$) has been combined with a minimum cluster threshold (e.g. 10-voxel contiguity) (Schmälzle et al., 2013). However, while such an option may be available with one group under one condition, it usually does not carry over to the contrast between two groups or conditions, as illustrated in our experimental results of group comparison (Fig. 3). The permutation test implemented in the ISC Toolbox through sign flipping between two condition is equivalent to EWP for one group; as shown here, the EWP approach violates the exchangeability assumption and leads to largely inflated FPR (top row, Fig. 2)

Advantages of nonparametric methods through subject-wise resampling

Among the nonparametric approaches examined here, the EW methods mostly achieve poor FPR controllability, and they should not be adopted in any ISC group analysis. In contrast, SWB works relatively well for both the one-group case (including the BGC subset $\boldsymbol{R}_{12}$) and group subsets (including each of $\boldsymbol{R}_{11}$ and $\boldsymbol{R}_{22}$ versus $\boldsymbol{R}_{12}$) although its FPR controllability is sensitive to the correlation magnitude ψ. Lastly, SWP is virtually ideal for group comparisons but performs poorly for the one-group scenario.

Within the subject-wise resampling methods there were interesting differences in performance observed, most likely due to the consequences of each distinct resampling mechanism on $\boldsymbol{R}^{(n)}$. In bootstrapping, there is a chance that a single subject is sampled more than once (as resampling is performed with replacement), which leads to the presence of artificial "1"s (i.e., perfect correlation, see last row in Table 1) in the resamplings. Such artificial values would introduce biases into the median estimation, as well as skewness toward the right tail in the resulting median distribution of the SWB. Even though the chosen centrality measure, the median, is less sensitive to such biases than the mean would be, still the occurrence of multiple artificial values results in overly conservative inferences when $\psi = 0$ (i.e., the case corresponding to all elements being fully independent of each other), regardless of the number of subjects. In addition, the biases lead to liberal inferences at the other extreme $\psi = 0.5$, though the impact here tends to asymptotically vanish as the number of subjects increases. On the other hand, the virtually ideal performance of SWP with two groups and its insensitivity to ψ may be due to the utilization of the between-group subset $\boldsymbol{R}_{12}$ as a leverage, while the lack of such a leverage in the case with one group is likely the cause for the unacceptable performance in terms of FPR controllability when $\psi \neq 0$.

There are several benefits to the nonparametric methods for ISC analysis proposed here. Unlike conventional parametric inferences such as the Student's *t*-test, ANOVA and GLM, the new nonparametric approaches bear fewer assumptions about the data structure at the group level, and they also have no need of any prior information about the time series at the individual level. For example, the Gaussianity assumption is not required. They are also less sensitive (and possibly immune) to data anomalies such as skewed distribution, outliers, etc. Furthermore, nonparametric methods also allow the option to include additional family-wise error control (Eklund et al., 2015) through approaches such as threshold-free cluster enhancement (Smith and Nichols, 2008).

The new nonparametric methods proposed here have further advantages over existing approaches. For example, they require significantly less computational time, as they resample ISCs rather than reshuffling the time series themselves. Additionally, the proposed nonparametric tests are executed directly on the ISC values $\{r_{ij}, i > j\}$, without the need of *z*-transformation, leading to less memory demand and lower computation cost than the previous approaches through resampling at the time series level. Additionally, bootstrapping on *z*-values may pick up some diagonals in (2), leading to a difficult situation with indeterminate values that are avoided when using *z*-values. Using the median as a centrality measure instead of averaging on the Fisher's *z*-values avoids further potential pitfalls of mean values if data are skewed or lopsided. The proposed approaches here also allow us to directly explore and evaluate the FPR controllability with the patterned correlation structure $\boldsymbol{P}^{(n)}$, unlike others that randomize the time series at the individual level, where it would be difficult to generate realistic ISC data of $\{r_{ij}\}$ (or its associated $\{z_{ij}\}$) that bear the correlation structure $\boldsymbol{P}^{(n)}$. Furthermore, to our best knowledge, the following approach of ours to group ISC and group comparisons have not been explored in the literature: 1) group analysis for BGC, $\boldsymbol{R}_{21}$, through SWB, 2) the comparison of ISCs through SWP between two groups ($\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{22}$), and 3) the comparison of WGC and BGC ($\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{21}$, or $\boldsymbol{R}_{22}$ vs $\boldsymbol{R}_{21}$) through SWP.

Importantly, the presented methodology with SW resampling, as opposed to the EW version, keeps the integrity of the variance-covariance structure as well the independence assumption. In contrast, sign flipping at the element level, for example as implemented for comparing two conditions in the ISC Toolbox, would break the patterned correlation structure in $\boldsymbol{P}^{(n)}$, leading to negative variances or a variance-covariance matrix that is not PSD (mathematically, SWP keeps constant eigenvalues of the correlation matrix $\boldsymbol{P}^{(n)}$ while

EWP does not). Another problematic aspect of EW sign flipping is that the difference of two ISC values is not inherently interpretable (e.g., it may lie beyond [−1, 1]).

Limitations of nonparametric approaches

With the advantage of having a parsimonious distribution assumption also come a few shortcomings and limitations. 1) None of the proposed methods work for all scenarios. Specifically, even though SWP shows near perfect FPR control with the two-sample case, its FPR performance is very poor for one group. On the other hand, SWB provides a reasonable FPR performance among alternatives, but it does not work well with two groups. 2) Although SWB showed the best performance among the methods tested for one group, it leaves some room for better alternatives. For example, it is too conservative in terms of FPR when ψ is quite close to 0, probably because bundling each subject's elements distorts the null distribution when those within-subject elements are virtually independent with each other. It is also too liberal when ψ is close to 0.5 (though specificity increases with sample size). 3) Even though the nonparametric approaches explored here are less computationally intensive than those implemented at the EPI time series level, the computational cost is still high compared to conventional parametric methods. 4) The methodology is not flexible in the sense that each scenario (e.g., one or more groups, inclusion of explanatory variables) would have to be dealt with separately. 5) The significance in a nonparametric test has a ceiling value limited by the number of resampling. For example, 5000 realizations of permutations or bootstrapping would lead to the lowest voxel-wise p-value down to 2×10^{-4}. 6) Little information could be pulled out from the nonparametric analyses other than simple inferences. For example, these approaches cannot offer a specific estimate or interpretation of the relatedness parameter ψ (e.g., what does it mean in terms of data variability when $\psi = 0.5$? Is ψ homogeneous across the brain?). 7) It is well known that serial correlation intrinsically exists in the residuals or unaccounted-for part of the time series model for FMRI data due to physiological (cardiac and respiratory) confounds and thermal fluctuations in the scaner, which may lead to biased ISC estimates when heterogeneity of serial correlation occurs across subjects or brain regions (Arbabshirani et al., 2014). Nevertheless, recent investigation indicates that the impact of biased estimations on statistical inferences is minimal or even negligible (Arbabshirani et al., 2014).

In light of these limitations, we will address several such issues in Part II, exploring the possibility of employing some new parametric approaches to the ISC data structure that provide a finer characterization, through proper parameterization and effects partitioning.

Conclusion

By framing the ISC group analysis with the specific variance-covariance structure among the elements, we have been able to lay out a patterned correlation matrix $\boldsymbol{P}^{(n)}$ among the ISC values, and to rigorously examine the FPR controllability and power achievement among a variety of methods through simulations. Through subject-wise instead of element-wise resampling of the ISC values, we found that 1) subject-wise permutation (SWP) testing is an ideal approach for handling the comparison between two within-group ISC components, between one within-group ISC and the between-group ISC component, and 2) subject-wise bootstrapping (SWB) is the best choice when inferring ISC for one group. An open source AFNI program, 3dNPT, is available to perform these approaches (http://afni.nimh.nih.gov).

Acknowledgments

Our work benefited significantly from the statistical computational language and environment *R*, its many packages, and the great support of the *R* community. All the plots were created in *R* with the base graphics library. We gratefully acknowledge the advice of J.-P. Kauppi, J. Tohka, and J. Pajula for using the ISC Toolbox. The research and writing of the paper were supported by the NIMH and NINDS Intramural Research Programs (grant number: 1ZICMH002888) of the NIH/HHS, USA, and by a grant (2012-0230) from the Asan Institute for Life Sciences, Asan Medical Center, Seoul, Korea. This work utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov).

References

Abrams DA, Ryali S, Chen T, Chordia P, Khouzam A, Levitin DJ, Menon V. Inter-subject synchronization of brain responses during natural music listening. The European Journal of Neuroscience. 2013; 37(9):1458–69. [PubMed: 23578016]

Anderson MJ. A new method for non-parametric multivariate analysis of variance. Aust Ecol. 2001; 26:32–46.

Arbabshirani MR, Damaraju E, Phlypo R, Plis S, Allen E, Ma S, Mathalon D, Preda A, Vaidya JG, Adali T, Calhoun VD. Impact of autocorrelation on functional connectivity. NeuroImage. 2014; 102:294–308. [PubMed: 25072392]

Barbé K, Van Moer W, Nagels G. Fractional-Order Time Series Models for Extracting the Haemodynamic Response From Functional Magnetic Resonance Imaging Data. IEEE Transactions on Biomedical Engineering. 2012; 59(8):2264–2272. [PubMed: 22677309]

Bartels A, Zeki S. The chronoarchitecture of the human brain - natural viewing conditions reveal a time-based anatomy of the brain. NeuroImage. 2004; 22:419–433. [PubMed: 15110035]

Bridwell DA, Roth C, Gupta CN, Calhoun VD. Cortical Response Similarities Predict which Audiovisual Clips Individuals Viewed, but Are Unrelated to Clip Preference. PloS one. 2015; 10(6):e0128833–e0128833. [PubMed: 26030422]

Buxton RB, Uludağ K, Dubowitz DJ, Liu TT. Modeling the hemodynamic response to brain activation. Neuroimage. 2004; 23(Suppl 1):S220–33. [PubMed: 15501093]

Cantlon JF, Li R. Neural activity during natural viewing of Sesame Street statistically predicts test scores in early childhood. PLoS Biol. 2013; 11(1):e1001462.doi: 10.1371/journal.pbio.1001462 [PubMed: 23300385]

Chen G, Saad ZS, Adleman NE, Leibenluft E, Cox RW. Detecting the subtle shape differences in hemodynamic responses at the group level. Front Neurosci. 2015; 9:375.doi: 10.3389/fnins.2015.00375 [PubMed: 26578853]

Cohen MS. Parametric analysis of fMRI data using linear systems methods. NeuroImage. 1997; 6(2):93–103. [PubMed: 9299383]

Cox RW. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res. 1996; 29:162–173. http://afni.nimh.nih.gov/afni. [PubMed: 8812068]

Eklund, A.; Nichols, T.; Knutsson, H. Can parametric statistical methods be trusted for fMRI based group studies?. 2015. http://arxiv.org/abs/1511.01863

Friston K, Fletcher P, Josephs O, Holmes A, Rugg M, Turner R. Event-related fMRI: characterizing differential responses. NeuroImage. 1998; 7(1):30–40. [PubMed: 9500830]

Gower JC, Krzanowski WJ. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. J R Stat Soc: Ser C: Appl Stat. 1999; 48:505–519.

Hasson U, Nir Y, Levy I, Fuhrmann G, Malach R. Intersubject synchronization of cortical activity during natural vision. Science. 2004; 303:1634–1640. [PubMed: 15016991]

Hasson U, Furman O, Clark D, Dudai Y, Davachi L. Enhanced intersubject correlations during movie viewing correlate with successful episodic encoding. Neuron. 2008a; 57:452–462. [PubMed: 18255037]

Hasson U, Landesman O, Knappmeyer B, Vallines I, Rubin N, Heeger DJ. Neurocinematics: The Neuroscience of Film. Projections Vol. 2008b; 2(1):1–26.

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N. A hierarchy of temporal receptive windows in human cortex. Journal of Neuroscience. 2008; 28(10):2539–2550. [PubMed: 18322098]

Hasson U, Malach R, Heeger DJ. Reliability of cortical activity during natural stimulation. Trends in Cognitive Science. 2010; 14(1):40–48.

Herbec A, Kauppi JP, Jola C, Tohka J, Pollick FE. Differences in fMRI intersubject correlation while viewing unedited and edited videos of dance performance. Cortex. 2015; 71:341–348. [PubMed: 26298503]

Honey CJ, Thomson CR, Lerner Y, Hasson U. Not lost in translation: Neural responses shared across languages. Journal of Neuroscience. 2012; 32(44):15277–15283. [PubMed: 23115166]

Jo HJ, Saad ZS, Simmons WK, Milbury LA, Cox RW. Mapping sources of correlation in resting state FMRI, with artifact detection and removal. Neuroimage. 2010; 52(2):571–82. [PubMed: 20420926]

Jola C, McAleer P, Grosbras MH, Love SA, Morison G, Pollick FE. Uni-and multisensory brain areas are synchronised across spectators when watching unedited dance recordings. i-Perception. 2013; 4(4):265–284. [PubMed: 24349687]

Kauppi JP, Jääskeläinen IP, Sams M, Tohka J. Inter-Subject Correlation of Brain Hemodynamic Responses During Watching a Movie: Localization in Space and Frequency. Front Neuroinformatics. 2010; 4:5.

Kauppi JP, Pajula J, Tohka J. A Versatile Software Package for Inter-subject Correlation Based Analyses of fMRI. Frontiers in Neuroinformatics. 2014; 8:2. [PubMed: 24550818]

Mantini D, Hasson U, Betti V, Perrucci MG, Romani GL, Corbetta M, Orban GA, Vanduffel W. Interspecies activity correlations reveal functional correspondence between monkey and human brain areas. Master thesis, Aalto University, Finland. Nat Methods. 2012; 9:277–282. [PubMed: 22306809]

Mériaux S, Roche A, Dehaene-Lambertz G, Thirion B, Poline JB. Combined permutation test and mixed-effect model for group average analysis in fMRI. Hum Brain Mapp. 2006; 27(5):402–410. [PubMed: 16596617]

Nichols TE, Holmes AP. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum Brain Mapp. 2002; 15(1):1–25. [PubMed: 11747097]

Nummenmaa, Glerean LE, Viinikainen M, Jääskeläinen IP, Hari R, Sams M. Emotions promote social interaction by synchronizing brain activity across individuals. Proc Natl Acad Sci U S A. 2012; 109(24):9599–9604. [PubMed: 22623534]

Pajula J, Tohka J. Effects of spatial smoothing on inter-subject correlation based analysis of FMRI. Magnetic Resonance Imaging Vol. 2014; 32(9):1114–1124.

Potes C, Brunner P, Gunduz A, Knight RT, Schalk G. Spatial and temporal relationships of electrocorticographic alpha and gamma activity during auditory processing. NeuroImage. 2014; 97:188–195. [PubMed: 24768933]

Reiss PT, Stevens MHH, Shehzad Z, Petkova E, Milham MP. On distance- based permutation tests for between-group comparisons. Biometrics. 2010; 66:636–643. [PubMed: 19673867]

Schmälzle R, Häcker F, Renner B, Honey C, Schupp H. Neural correlates of risk perception during real-life risk communication. J Neurosci. 2013; 33:10340–10347. [PubMed: 23785147]

Schmälzle R, Häcker F, Honey C, Schupp H, Hasson U. Engaged listeners: shared neural processing of powerful political speeches. Soc Cogn Affect Neurosci. 2015; 10(8):1137–1143. [PubMed: 25653012]

Seber, GAF. A Matrix Handbook for Statisticians. John Wiley & Sons, Inc; 2008.

Shehzad Z, Kelly, Reiss PT, Craddock RC, Emerson JW, McMahon K, Copland DA, Castellanos FX, Milham MP. A multivariate distance-based analytic framework for connectome-wide association studies. Neuroimage. 2014; 93:74–94. [PubMed: 24583255]

Silver NC, Dunlap WP. Averaging correlation coefficients: Should Fisher's z transformation be used? Journal of Applied Psychology. 1987; 72(1):146–148.

Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage. 2008; 44:83–98. [PubMed: 18501637]

Thiede, A. Master thesis. Aalto University; Finland: 2014. Magnetoencephalographic (MEG) Inter-subject Correlation using Continuous Music Stimuli.

Trost W, Sascha Frühholz S, Cochrane T, Cojan1 Yann, Vuilleumier P. Temporal dynamics of musical emotions examined through intersubject synchrony of brain activity. Soc Cogn Affect Neurosci. 2015; doi: 10.1093/scan/nsv060

Wilson SM, Molnar-Szakacs I, Iacoboni M. Beyond Superior Temporal Cortex: Intersubject Correlations in Narrative Speech Comprehension. Cereb Cortex. 2008; 18(1):230–242. [PubMed: 17504783]

Winkler AM, Ridgway GR, Webster MA, Smith SM, Nichols TE. Permutation inference for the general linear model. Neuroimage. 2014; 92:381–397. [PubMed: 24530839]

Winkler AM, Webster MA, Vidaurre D, Nichols TE, Smith SM. Multi-level block permutation. Neuroimage. 2015; 123:253–268. [PubMed: 26074200]

Zapala MA, Schork NJ. Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. Proc Natl Acad Sci. 2006; 103:19430–19435. [PubMed: 17146048]

Appendix A. List of acronyms used in the paper

ANOVA	analysis of variances
BGC	between-group correlation
EW	element-wise
EWB	element-wise bootstrapping
EWP	element-wise permutations
FPR	false positive rate
GLM	general linear model
HDR	hemodynamic response
ISC	inter-subject correlation
PSD	positive semi-definite
SW	subject-wise
SWB	subject-wise bootstrapping
SWP	subject-wise permutations
WGC	within-group correlation

Appendix B. Necessity and sufficiency for the positive semi-definiteness of the matrix P(n)

Here, we demonstrate and prove an allowed interval of ψ values for the matrix $\boldsymbol{P}^{(n)}$ to be positive semidefinite (PSD) for any $n \geq 2$. We note the following four important properties for a real symmetric matrix $\boldsymbol{A}$ (e.g., Seber, 2008)):

A. $\boldsymbol{A}$ is PSD if and only if all of its eigenvalues are non-negative ($\lambda_i \geq 0$);

B. If the matrix $\boldsymbol{A}$ is PSD, then all its principal submatrices are PSD;

C. If $\boldsymbol{A}$ is diagonally dominant with non-negative diagonal entries (i.e., within each row, the magnitude of the non-negative diagonal element is greater than or equal to the sum of the magnitudes of all off-diagonal elements), then $\boldsymbol{A}$ is PSD.

D. If $\boldsymbol{A}$ is a correlation matrix, then $\boldsymbol{A}$ is PSD.

First, the PSD property holds trivially when the number of subjects is $n = 2$, as $\boldsymbol{P}^{(2)} = [1]$ and is independent of ψ. For $n = 3$, there are $N = 3$ z-values (z_{21}, z_{31} and z_{32}), and their correlation matrix is,

$$\boldsymbol{P}^{(3)} = \begin{array}{c} \\ z_{21} \\ z_{31} \\ z_{32} \end{array} \left(\begin{array}{ccc} z_{21} & z_{31} & z_{32} \\ 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{array} \right).$$

The characteristic polynomial for this matrix is $f(\lambda) = (\lambda - 2\psi - 1)(\lambda^2 + \lambda + 1)$, which can be solved to provide a necessary condition on ψ for having only positive eigenvalues λ: $-0.5 \leq \psi \leq 1$. By Property A, $\boldsymbol{P}^{(3)}$ is PSD for this interval of ψ.

We note that matrix rows and columns can be rearranged without changing the eigenvalues. Therefore, for any $n \geq 4$ (with $N \geq 6$), we choose to start the list of columns and rows of $\boldsymbol{P}^{(n \geq 4)}$ with the values z_{31}, z_{41}, z_{32} and z_{42}, so that the 4×4 principal submatrix is always

$$\boldsymbol{P}_0^{(n \geq 4)} = \begin{array}{c} \\ z_{31} \\ z_{41} \\ z_{32} \\ z_{42} \end{array} \left(\begin{array}{cccc} z_{31} & z_{41} & z_{32} & z_{42} \\ 1 & \rho & \rho & 0 \\ \rho & 1 & 0 & \rho \\ \rho & 0 & 1 & \rho \\ 0 & \rho & \rho & 1 \end{array} \right).$$

This submatrix has the characteristic polynomial $f(\lambda) = (\lambda - 1)^2(\lambda + 2\psi - 1)(\lambda - 1 - 2\psi)$, for which $\lambda > 0$ ($\Rightarrow \boldsymbol{P}_0^{(n \geq 4)}$ is PSD, by Property A) when: $-0.5 \leq \psi \leq 0.5$. Due to Property B, this provides a condition that must also be met for the full $\boldsymbol{P}^{(n)}$. We further investigate the positive and negative subintervals of this bounding range separately.

We first show that $0 \leq \psi \leq 0.5$ is a sufficient condition for $\boldsymbol{P}^{(n\geq 4)}$ to be PSD. Since the set of PSD symmetric matrices is known to be convex, we investigate the boundary values of ψ in this positive subinterval. When $\psi = 0$, then $\boldsymbol{P}^{(n)}$ is simply the identity matrix, which is PSD. For $\psi = 0.5$, we note that we can recognize the structure of $\boldsymbol{P}^{(n)}$ is equivalent to a matrix derived in the following way. Let $\{X_i, i = 1, 2, \ldots n\}$ be independent and identically distributed with equal probability of taking the value 1 or −1. Then define the variable $Y_{ij}=(X_i+X_j)/\sqrt{2}, 1 \leq j < i \leq n$. If we flatten the $N = n(n-1)/2$ values of $\{Y_{ij}, 1 \leq j < i \leq n\}$, the correlation matrix of the resulting vector will have the same structure as $\boldsymbol{P}^{(n\geq 4)}$ with $\psi =$ 0.5. Any correlation matrix of real random variables will be PSD (e.g., by Property D), thereby proving the property for $\boldsymbol{P}^{(n\geq 4)}$ with $\psi = 0.5$.

Lastly, we investigate where the negative subinterval of ψ values provides a PSD $\boldsymbol{P}^{(n\geq 4)}$. Since $\psi = 0$ has already been shown to produce a PSD matrix, we search for a lower bound of ψ. By Property C, $\boldsymbol{P}^{(n)}$ will be PSD if it is diagonally dominant, which can be explored in general by counting the number of elements per row/column having a given value. The set of $\{z_{ij}, i > j\}$ has a cardinality of $N = n(n-1)/2$, so that each row or column of $\boldsymbol{P}^{(n)}$ has length N. Of this number, for any allowed pair of indices $\{i_1, j_1\}$, there are $n - 2$ sets of allowed $\{i_2, j_2\}$ containing exclusively i_1 or j_1 (that is, where the cardinality of $\{i_1, j_1\} \cap \{i_2, j_2\}$ is 1; see the definition of $\boldsymbol{P}^{(n)}$ in (4) of the main text); hence, there are exactly $2(n-2)$ elements with value ψ in any row or column of $\boldsymbol{P}^{(n)}$. The diagonal element is always unity, and the remaining $N - 2(n-1) - 1 = (n-2)(n-3)/2$ elements in a row or column are 0. Therefore, $\boldsymbol{P}^{(n)}$ is diagonally dominant for negative ψ wherever $1 \geq 2(n-2)\psi$, or where $-1/[2(n-2)] \leq \psi < 0$. Regarding further negative ψ values within the original subinterval, we note that the matrix $\boldsymbol{P}^{(n)}$ has an eigenvector of 1_N associated with eigenvalue $\lambda = 1+2(n-2)\psi$, as $\boldsymbol{P}^{(n)}1_N = [1+2(n-2)\psi]1_N$. For values of $\psi < -1/[2(n-2)]$, this eigenvalue becomes negative, so that $\boldsymbol{P}^{(n)}$ is not PSD for these ψ values.

In summary, the sufficient and necessary condition for the PSD property of $\boldsymbol{P}^{(n)}$ can be written as:

$$\boldsymbol{P}^{(n)}\,\mathrm{is\,PSD\,if}\,\rho \in \begin{cases} [-0.5, 1], & n=3; \\ \left[-\frac{1}{2(n-2)}, 0.5\right], & n \geq 4. \end{cases}$$

Appendix C. Results for the novel indirect group comparisons

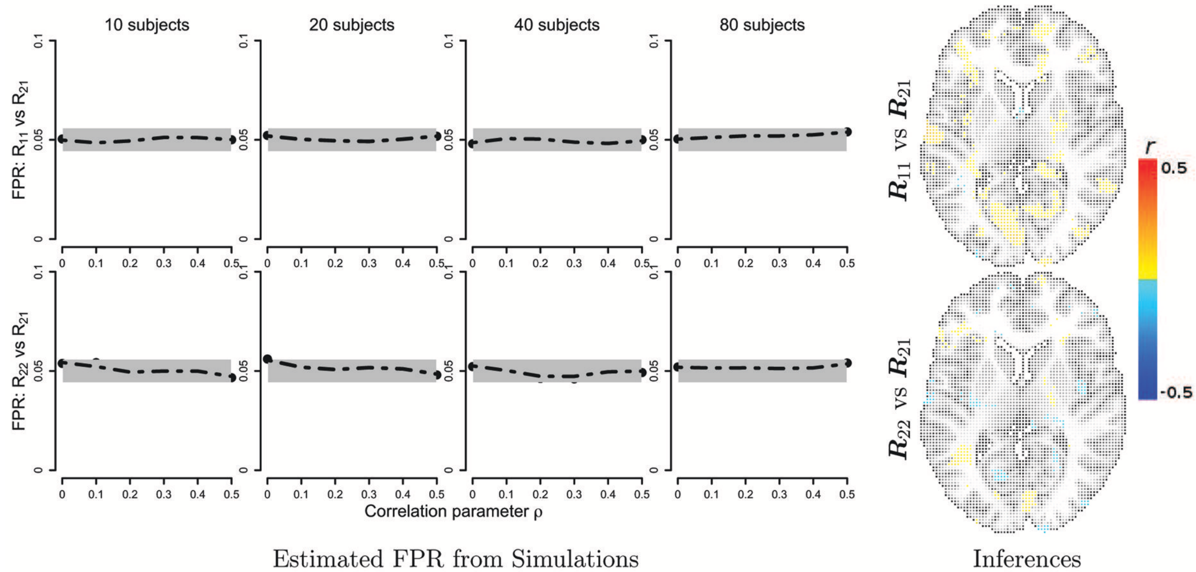


Figure 4.

Performances of subject-wise permutations (SWP) on the comparisons of within-group versus between-group ISCs (i.e., indirect comparisons) are shown in terms of FPR controllability (left) and inferences (right) with an experimental dataset. The simulation parameters and experimental dataset are the same as in Fig. 2 and Fig. 3, respectively, but the scale for FPR (left) is different from Fig. 2. The gray line of FPR = 0.05 indicates the 95% confidence band of the target (or nominal) value (with a band width of 0.012 for each simulation with 5000 realizations).

Highlights

- Inter-subject correlation shows similarities of response to naturalistic conditions
- A patterned matrix is formulated to characterize the relationship among ISC values
- Previous methods of ISC group analysis are problematic in controlling for FPR
- Subject-wise permutation testing is ideal for comparing two groups of ISC
- Bootstrapping is the best nonparametric method to make ISC inferences for one group

$$\boldsymbol{R}^{(n)} =$$

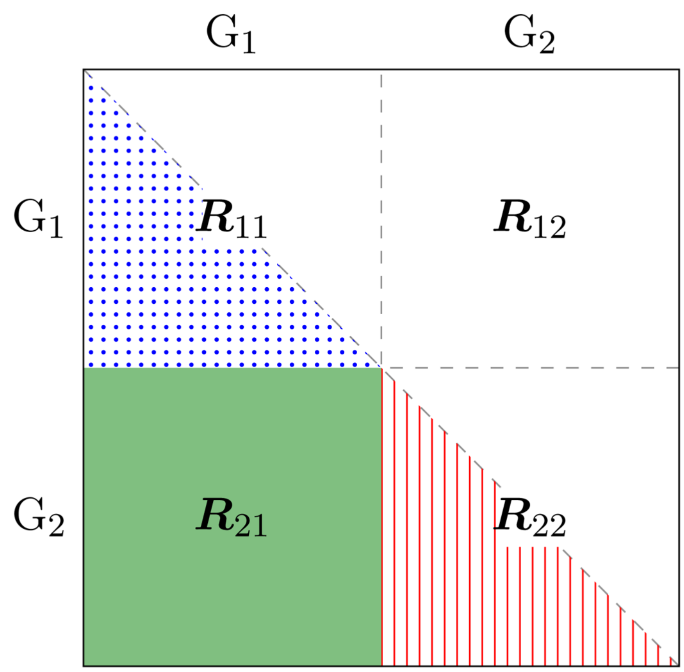


Figure 1.

Schematic illustration of ISC data $\boldsymbol{R}^{(n)}$ from two groups. With n_1 and n_2 subjects, respectively, in the two groups G_1 and G_2, the $N_1 = \frac{1}{2}n_1(n_1-1)$ elements in $\boldsymbol{R}_{11}$ (blue dots) and $N_2 = \frac{1}{2}n_2(n_2-1)$ elements in $\boldsymbol{R}_{22}$ (red lines) are WGC values while the $N_{12} = n_1 n_2$ elements in R_{21} (green) show the BGC values. Three meaningful comparisons can be formulated: $\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{22}$, $\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{21}$, and $\boldsymbol{R}_{22}$ vs $\boldsymbol{R}_{21}$, as discussed in the text.

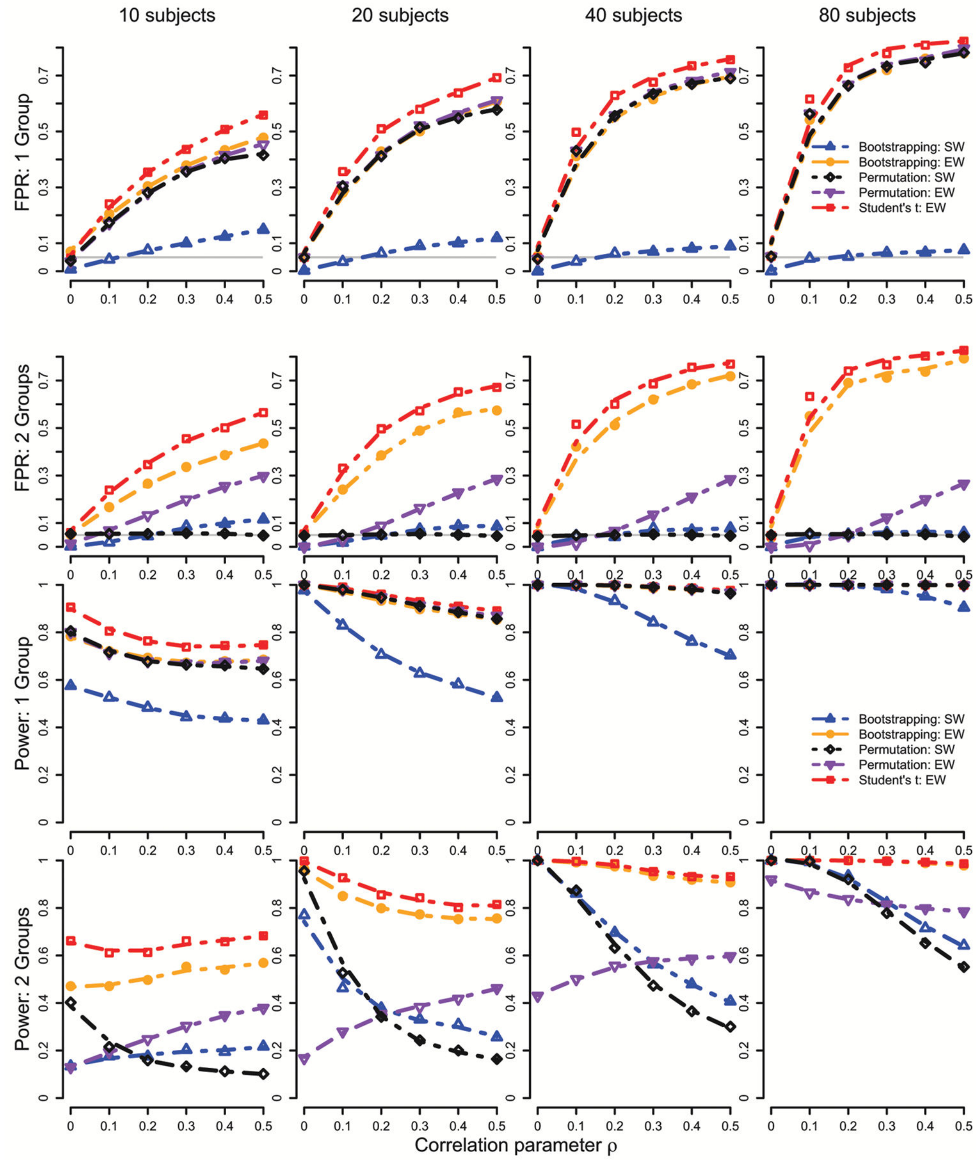
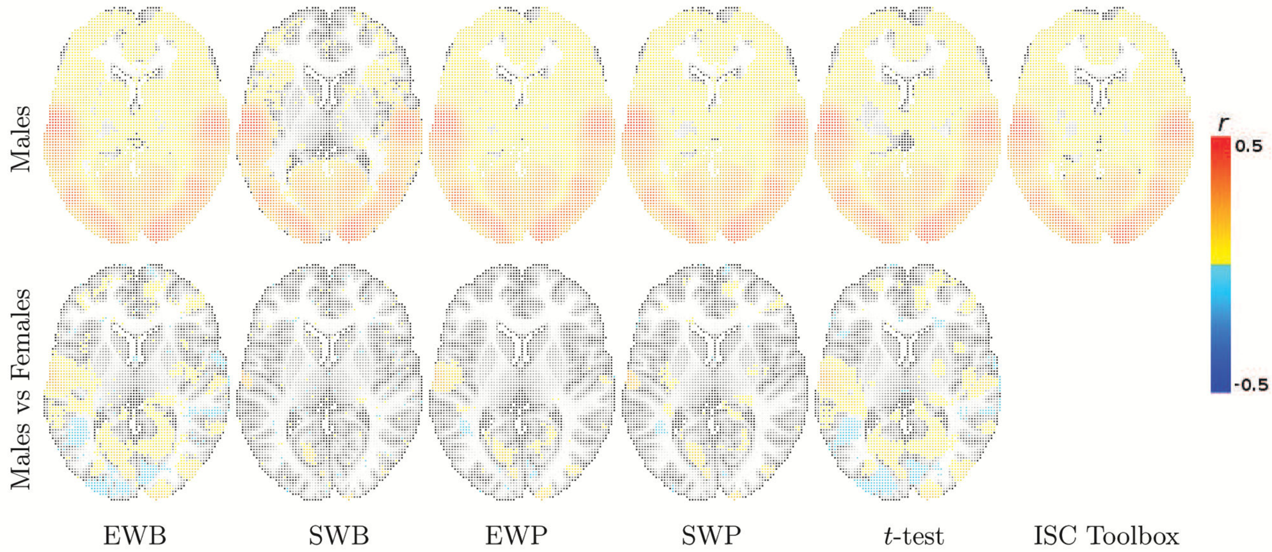


Figure 2.

Simulation parameters and results are shown here for five methods: SWP, EWB, SWP, EWP, and Student's *t*. False positive rate (FPR) performances are illustrated in the first two rows, and power achievement in the last two rows. Each of the four columns represents the number of subjects in each group (one group, $n = 10, 20, 40, 80$; two groups, $n_1 = n_2 = 10, 20, 40, 80$). The gray band of FPR = 0.05 in the first two rows indicates the 95% confidence interval of the target (or nominal) value (with a width of 0.012 for each simulation with 5000 realizations†), which is barely visible for the case of two groups (second row) due to its overlap with the results of SWP. The curves for FPR were fitted to the simulation results (plotting symbols) through a cubic smoothing spline. Among the three possible comparisons for the two-group scenario, only the direct contrast between the two WGCs, $\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{22}$, is shown here, but the results for other two indirect contrasts ($\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{21}$ and $\boldsymbol{R}_{22}$ vs $\boldsymbol{R}_{21}$) with SWP were similar (see Fig. 4 in Appendix C). The SWP testing is uniformly well-behaved and essentially ideal for two groups (black in the second row). On the other hand, in the one-group case, SWB offers a better compromise than any of the nonparametric alternatives considered here, particularly as n increases, even though it can be a little liberal or over-conservative (blue in the first row) depending on the amplitude of correlation ($0 \leq \psi \leq 0.5$). The comparisons of FPR and power among the five methods with the BGC component $\boldsymbol{R}_{21}$ (not shown here) are similar to the one group scenario (first row).

†The confidence band is computed with the assumption of a binomial distribution B(n, p), where $n = 5000$, p = PFR = 0.05.

(A) Visual Comparisons of One Slice



(B) Number of Suprathreshold Voxels at Various Significance Levels

Significance Level	One Group: Males						Group Comparison: Males vs Females				
	EWB	SWB	EWP	SWP	t-test	ISC Toolbox	EWB	SWB	EWP	SWP	t-test
0.05	153,298	116,068	158,911	158,164	152,822	159,836	48,946	7,690	6,877	13,410	46,587
0.01	144,967	91,572	149,789	149,792	143,248	148,481	31,524	1,610	2,864	3,268	28,145
0.005	141,745	82,074	148,939	146,242	139,489	$*$	26,647	814	2,142	1,635	23,109
0.001	134,979	64,778	143,157	138,154	131,609	134,514	19,046	197	1,242	378	15,330
0.0005	132,969	60,399	141,513	135,578	128,463	$*$	17,432	126	1,076	255	13,031

Figure 3.

Performance comparisons with an experimental dataset. (A) Axial views ($Z = 5$ mm; radiological convention: left is right) of ISC group results (thresholded by p-values, below) of an experimental dataset are illustrated for the five methods as well as the approach implemented in the ISC Toolbox. The colors code for the magnitude of correlation coefficients. For both one- and two-sample tests, the results are consistent with their FPR controllability. Specifically, for the male group ($n = 24$, upper panel, two-tailed significance level $p = 0.001$) all the other five methods are more liberal than SWB. Although not visually obvious in the color coding, the group ISC estimates through averaging across subjects in the ISC Toolbox tend to be biased relative to the medians from the other four nonparametric methods. For two-group comparison ($n = 48$, lower panel, two-tailed significance level $p = 0.05$; group comparison testing currently not available in the ISC Toolbox), EWP was much more liberal relative to SWP for some regions, while being over-conservative for others; EWB was too liberal, SWB tended to be slightly more conservative, and both rendered noisier results; and Student's t-test was the worst. The SWP performance for the other two indirect contrasts ($\boldsymbol{R}_{11}$ vs $\boldsymbol{R}_{21}$ and $\boldsymbol{R}_{22}$ vs $\boldsymbol{R}_{21}$) are shown in Fig. 4 of Appendix C. We note that 1) multiple testing correction was not performed so that voxel-wise comparisons among the methods could be directly visualized; and 2) except for the ISC Toolbox, which adopts direct averaging across ISC values, all other methods rendered virtually the same group estimate for ISC, but differed in significance detection (i.e., the color at each voxel is roughly the same across the first five testing methods if the significance survives the corresponding threshold). The performance comparisons among the five methods with the BGC component $\boldsymbol{R}_{21}$ (not shown here) are similar to the one group scenario for males (first row), $\boldsymbol{R}_{11}$. (B) List of voxels that pass five significance levels. The default setting in ISC Toolbox does not provide thresholding at voxel-wise significance levels of 0.005 and 0.0005 (marked with * in the table).

Table 1

Illustrative comparisons among nonparametric methods (permutation and bootstrapping, each at the element- and subject-wise level) for an example of $\boldsymbol{R}^{(6)}$. All samples, sign flips and group reassignments, are randomly chosen. For EWB and EWP, the matrix structure becomes broken by the randomization and no longer matters. The randomization process for the BGC subset $\boldsymbol{R}_{21}$ (not demonstrated here), even though formulated in the case of two groups, is similar to the scenario with one group.

	One Group	Two Groups
$\boldsymbol{R}^{(6)}$	$\begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \left(\begin{array}{cccccc} 1 & r_{12} & r_{13} & r_{14} & r_{15} & r_{16} \\ r_{21} & 1 & r_{23} & r_{24} & r_{25} & r_{26} \\ r_{31} & r_{32} & 1 & r_{34} & r_{35} & r_{36} \\ r_{41} & r_{42} & r_{43} & 1 & r_{45} & r_{46} \\ r_{51} & r_{52} & r_{53} & r_{54} & 1 & r_{56} \\ r_{61} & r_{62} & r_{63} & r_{64} & r_{65} & 1 \end{array}\right) \end{array}$	$\begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \left(\begin{array}{cccccc} 1 & r_{12} & r_{13} & r_{14} & r_{15} & r_{16} \\ r_{21} & 1 & r_{23} & r_{24} & r_{25} & r_{26} \\ r_{31} & r_{32} & 1 & r_{34} & r_{35} & r_{36} \\ \hline r_{41} & r_{42} & r_{43} & 1 & r_{45} & r_{46} \\ r_{51} & r_{52} & r_{53} & r_{54} & 1 & r_{56} \\ r_{61} & r_{62} & r_{63} & r_{64} & r_{65} & 1 \end{array}\right) \end{array}$
EWP	Flipped sign: $r_{21}, r_{51}, r_{61}, r_{32}, r_{62}, r_{63}, r_{54}$ $\begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \left(\begin{array}{cccccc} & & & & & \\ -r_{21} & & & & & \\ r_{31} & -r_{32} & & & & \\ r_{41} & r_{42} & r_{43} & & & \\ -r_{51} & r_{52} & r_{53} & -r_{54} & & \\ -r_{61} & -r_{62} & -r_{63} & r_{64} & r_{65} & \end{array}\right) \end{array}$	Reassigned correlation coefficients G1: r_{21}, r_{54}, r_{64}; G2: r_{31}, r_{32}, r_{64}
SWP	Flipped sign: S_1, S_4, S_6 $\begin{array}{cc} & \begin{array}{cccccc} S_1 & S_2 & S_3 & S_4 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \\ S_6 \end{array} & \left(\begin{array}{cccccc} & & & & & \\ -r_{21} & & & & & \\ -r_{31} & r_{32} & & & & \\ r_{41} & -r_{42} & -r_{43} & & & \\ -r_{51} & r_{52} & r_{53} & -r_{54} & & \\ r_{61} & -r_{62} & -r_{63} & r_{64} & -r_{65} & \end{array}\right) \end{array}$	Reassigned G1: S_2, S_5, S_6; G2: S_1, S_3, S_4 $\begin{array}{cc} & \begin{array}{cccccc} S_2 & S_5 & S_6 & S_1 & S_3 & S_4 \end{array} \\ \begin{array}{c} S_2 \\ S_5 \\ S_6 \\ S_1 \\ S_3 \\ S_4 \end{array} & \left(\begin{array}{cccccc} & & & & & \\ r_{52} & & & & & \\ r_{62} & r_{65} & & & & \\ \hline r_{21} & r_{51} & r_{61} & & & \\ r_{32} & r_{53} & r_{63} & r_{31} & & \\ r_{42} & r_{54} & r_{64} & r_{41} & r_{43} & \end{array}\right) \end{array}$
EWB	Sampled correlation coefficients: $r_{21}, r_{21}, r_{32}, r_{41}, r_{43}, r_{43}, r_{52}, r_{53}$, $r_{53}, r_{53}, r_{45}, r_{61}, r_{63}, r_{64}, r_{64}$	Sampled correlation coefficients: G1: r_{21}, r_{32}, r_{32}; G2: r_{54}, r_{64}, r_{64}
SWB	Sampled subjects: $S_1, S_3, S_3, S_5, S_5, S_6$ $\begin{array}{cc} & \begin{array}{cccccc} S_1 & S_3 & S_3 & S_5 & S_5 & S_6 \end{array} \\ \begin{array}{c} S_1 \\ S_3 \\ S_3 \\ S_5 \\ S_5 \\ S_6 \end{array} & \left(\begin{array}{cccccc} & & & & & \\ r_{31} & & & & & \\ r_{31} & \boxed{1} & & & & \\ r_{51} & r_{53} & r_{53} & & & \\ r_{61} & r_{63} & r_{63} & r_{65} & & \\ r_{61} & r_{63} & r_{63} & r_{65} & \boxed{1} & \end{array}\right) \end{array}$	Sampled subjects G1: S_2, S_2, S_3; G2: S_4, S_6, S_6 $\begin{array}{cc} & \begin{array}{cccccc} S_2 & S_2 & S_3 & S_4 & S_6 & S_6 \end{array} \\ \begin{array}{c} S_2 \\ S_2 \\ S_3 \\ S_4 \\ S_6 \\ S_6 \end{array} & \left(\begin{array}{cccccc} & & & & & \\ \boxed{1} & & & & & \\ r_{32} & r_{32} & & & & \\ \hline r_{42} & r_{42} & r_{43} & & & \\ r_{62} & r_{62} & r_{63} & r_{64} & & \\ r_{62} & r_{62} & r_{63} & r_{64} & \boxed{1} & \end{array}\right) \end{array}$