Intraclass Correlations for Assessing Reliability in Presence of Within- and Between-Subject Variability in FMRI BOLD Responses

Alexander Yu Zhigalov
Neuroscience Center, University of Helsinki – a.zhigalov@gmail.com

Summit Suen
Institute of Statistical Science, Academia Sinica – summit.suen@gmail.com

Juin-Der Lee
College of Commerce, National Chengchi University – juinder.lee@gmail.com

Vincent S. C. Chien
Institute of Statistical Science, Academia Sinica – vincent@stat.sinica.edu.tw

Philip E.Cheng
Institute of Statistical Science, Academia Sinica – pcheng@stat.sinica.edu.tw

Michelle Liou*
Institute of Statistical Science, Academia Sinica – mliou@stat.sinica.edu.tw

Abstract

In this study, we consider a commonly used intraclass correlation (ICC) index in fMRI studies, and derive its asymptotic standard error. Recently, the fluctuation in fMRI BOLD responses was found to be a better predictor to a subject's age and task performance compared to the average responses. In a sense, the within- and between-subject variances of ICC values might be spatially distributed in a pattern informative about developmental or aging effects in a group of subjects having distinct demographic features. We also derive the ICC test statistics for assessing within- and between-subject reliability. Based on the test statistics, the supra-threshold voxels in an ICC map can be decided by controlling the false positive rate (FDR) using the voxel-FDR method based on the phase-randomization distribution or cluster-FDR method based on standard results from the random field theory. In the empirical study, the ICC statistics were applied to assessing the within- and between-subject reliability of BOLD responses in an 8-min eyes-closed/open task (4-min each) with 30 healthy subjects (15 males, average age 22.50±3.462). Because subjects received the acoustic instruction to close or open their eyes, the supra-threshold voxels in the ICC maps indicated that BOLD responses in the primary auditory and motor cortices were more reliable within individual subjects and between the functional states of eyes-closed and –open. The default mode regions, motor and occipital-parietal cortices, and thalamus were more reliable between subjects.

Keywords: asymptotic standard errors, BOLD signal, intraclass correlations, resting-state.

1. Introduction

The intraclass correlations (ICCs) assess the degree to which brain activation maps or connectivity networks resemble each other while subjects performing the same tasks twice or more. Theoretically, the ICC defines a ratio between two variances: one is the hypothetical true variance of outcomes across experimental replicates and the other is their observed variance. The true variance is measured without errors, and proportional to the observed variance. Although a variance ratio should lie between a range between zero and one, an observed ICC value can range from negative infinity to one because the true variance is estimated using empirical data. In test-retest functional MRI (fMRI) studies, ICC values were evaluated against either a Gaussian distribution by setting all negative values to zero, or a

complete-randomization distribution using the bootstrap method. In studies on group-level reliability in BOLD responses, the ICC values were evaluated using qualitative standards such as < 0.4 as poor, 0.4−0.75 as fair to good, and > 0.75 as excellent. Here we introduce an ICC statistic and derive its asymptotical standard error. The proposed statistic can be directly applied to pre-processed BOLD responses without making an assumption on the temporal composition of processes contributing to the responses. Because the proposed statistic can be expressed as a scalar-valued function of the empirical covariance matrix of BOLD time courses, the asymptotic distribution of standardized ICC values can be approximated by a Gaussian distribution, against which the voxel-wise ICC values can be evaluated. In the method section, we elaborate the step-by-step procedure for computing the statistics and specify the conditions to be satisfied for a proper application of the statistic to detect systematic responses to experimental interventions. In the empirical study, the ICC statistic was applied to assessing within- and between-subject reliability of BOLD responses in an 8-min eyes-closed/open task (4-min each) with 30 healthy subjects (15 males, average age 22.50±3.462). We made an empirical comparison between the phase-randomization test and the test based on the random field theory. We finally discuss applications of the ICC statistic in between-ROI, -subject and –center comparisons.

2. The ICC statistic

Given M experimental replicates (e.g., runs, subjects or centers), we denote $\underline{y}_i$ as the vector of pre-processed BOLD time course in the *i*th replicate, and $\boldsymbol{S}$ as the empirical covariance matrix across $\underline{y}_i$'s in a single voxel. We can compute $\underline{Y} = \underline{y}_1 + \cdots + \underline{y}_M$ as the global level data, and assume the global level data to be an additive sum of a true part plus the random error; that is, $\underline{Y} = \underline{T} + \underline{E}$, or equivalently, $\underline{y}_i = \underline{t}_i + \underline{e}_i$. The ICC statistic can be theoretically defined as the ratio between the variance of $\underline{T}$ over the variance of $\underline{Y}$, and allows contributions of individual $\underline{y}_i$ to $\underline{T}$ differently, that is,

$$\underline{t}_i = \lambda_i \underline{T} + c_i, \tag{1}$$

where λ_i is a positive constant with $\sum_i \lambda_i = 1$, and c_i is restricted to $\sum_i c_i = 0$. With the two restrictions, $\underline{T}$ can be equated to the sum of $\underline{t}_i$'s, and the variance of $\underline{t}_i$ is proportional to that of $\underline{T}$ by a constant λ_i^2. There are 2M+1 unknown parameters in (1) and M(M+1)/2 known elements, plus one restriction on λ_i, for estimating the parameters. In order to identify the parameters, M must be ≥ 3. The commonly used ICC statistic relies on the restriction $\lambda_i = 1/M$ for all replicates, which reduces the number of unknown parameters to M+1. This statistic introduces a scale-sensitive measure of reliability in that its value decreases as the variance of image intensity varies from one replicate to the other. However, the statistic is unaffected by adding an arbitrary constant to image intensity in any replicate. The ICC has the following form:

$$\widehat{\mathrm{ICC}} = \frac{M}{M-1}\left[1 - \mathrm{tr}(\boldsymbol{S})/(\underline{1}'\boldsymbol{S}\underline{1})\right] \tag{2}$$

where tr($\boldsymbol{S}$) denotes the trace of $\boldsymbol{S}$, and $\underline{1}$ is the summing vector of order M with the transposition $\underline{1}'$. The ICC statistic measures the reliability of $\underline{Y}$ when the model in (1) is true and $\lambda_i = 1/M$, but its empirical size falls within the range of (-∞, 1].

Let $\boldsymbol{S}$ have its population counterpart $\boldsymbol{\Sigma}$ and finite fourth moments, and vech$\boldsymbol{S}$ be the vector of non-duplicated elements in $\boldsymbol{S}$, with the identity vech$\boldsymbol{S}$ = K_M(vec$\boldsymbol{S}$), where K_M is of order M(M+1)/2 × M^2 and vec$\boldsymbol{S}$ denotes the vector of all elements in $\boldsymbol{S}$. As n is sufficiently large, it is known that $\sqrt{n}$(vech$\boldsymbol{S}$−vech$\boldsymbol{\Sigma}$) is asymptotically distributed according to the M(M+1)/2-dimensional Gaussian distribution with mean zero and the following covariance matrix,

$$2K_M(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})K_M'. \tag{3}$$

where $\otimes$ is the Kronecker product. The ICC statistic is a scalar-valued function of $\boldsymbol{S}$ and differentiable in a neighborhood of $\boldsymbol{S} = \boldsymbol{\Sigma}$. The variance of $\widehat{\mathrm{ICC}}$ can be estimated by

$$\mathrm{Var}(\widehat{\mathrm{ICC}}) = 2\mathrm{n}^{-1}\eta'\mathrm{K}_{\mathrm{M}}(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\mathrm{K}_{\mathrm{M}}'\eta, \tag{4}$$

where $\eta' = \partial\mathrm{ICC}/\partial\mathrm{vech}'\boldsymbol{\Sigma}$. The asymptotic distribution of $\widehat{\mathrm{ICC}}$ can be expressed as:

$$\sqrt{n}\left(\widehat{\mathrm{ICC}} - \mathrm{ICC}\right) \sim \mathrm{N}(0,\ 2\eta'\mathrm{K}_{\mathrm{M}}(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\mathrm{K}_{\mathrm{M}}'\eta). \tag{5}$$

As n is reasonably large, it is reasonable to assume that the standardized index,

$$Z = \left(\widehat{\mathrm{ICC}} - \mathrm{ICC}_0\right) / \sqrt{\mathrm{Var}(\widehat{\mathrm{ICC}})}\ , \tag{6}$$

is distributed as a standard Gaussian distribution with zero mean and unit variance given the hypothetical value of ICC_0. When $\mathrm{ICC}_0 = 0$, an empirical Z can be evaluated for statistical significance against a standard Gaussian distribution.

The asymptotic results in (3)-(6) are all derived by assuming that noise in the n image scans are independently and identically distributed for individual time courses. With temporally dependent errors, the asymptotic results may hold if noise in the n image scans are strictly stationary, and satisfy the following conditions: (a) the lagged correlation between $e_{i,j}$ and $e_{i,j+1}$ is uniformly bounded, with a bound less than 1 (where $e_{i,j}$ denotes the jth residuals in $\underline{y_i}$ after removing trends and experimental effects); (b) the maximum of lagged correlations between $e_{i,j}$ and $e_{i,j+k}$ goes to zero as k goes to infinity (Bradley, 2011). With temporally dependent $e_{i,j}$, the BOLD time courses were decomposed into different frequency bands to estimate reliability indices within each band. In studies involving long-term sensory stimulation, BOLD responses may exhibit long memory and other special types of dependence. The ICC statistic allows for a temporal model specified on noise, which is theoretically tractable when an ICC statistic can be expressed as a scalar-valued function of $\boldsymbol{S}$. An interested reader may refer to Dryden et al. (2010) for theoretical details.

3. Within- and between-subject reliability

In addition to regular pre-processing procedures (i.e., slice timing, and motion correction), the BOLD time courses in reliability assessment must be corrected for major trend effects in order to account for magnetic field drifts. After removing the trends, ICC can be directly applied to preprocessed time courses for assessing between-subject reliability, a case in which each subject is considered a replicate and M is the total number of subjects. The ICC statistic is computed using temporal information in each voxel, and the resulting Z score in (6) can be tested against a nominal Type-I error rate α in each individual voxel. The false discovery rate (FDR) control constrains the expected rate of a false classification of voxels into the reliable category at α. However, this holds when the voxel-wise *p*-values are independent and under a positive regression dependence, which exists when $e_{i,j}$ are distributed as Gaussian with non-negative correlation across voxels and the statistical tests are one-sided (Heller et al., 2006). For spatially coherent image data, it has been suggested to conduct peak- or cluster-FDR control over the false positive rate, which constrains the expected rate of a false classification of clusters into the reliable category at α. Because the number of clusters is much smaller than the number of voxels, the problem of multiple-hypothesis testing becomes less serious than the unconstrained procedure. There are also several peak- and cluster-FDR procedures available in the literature. For instance, the cluster-FDR procedure proposed in Chumbley et al. (2010) computes the threshold based on the uncorrected *p*-value for each cluster in the SPM. The uncorrected *p*-value for clusters above some thresholds can be computed using standard results from the random field theory (which accounts for spatial dependence in the data as captured by the maximum of a random field), and submitted to a FDR-control algorithm, which returns a threshold that controls the expected false-

discovery rate. In reliability assessment, we are interested in positive ICC values because the statistics estimate the ratio between the true and observed variances. In fMRI applications, there would be a large number of Z scores distributed below zero. In the empirical study, we compared the results between voxel- and cluster-FDR procedures. The p-values in the voxel-FDR were evaluated against the phase-randomization distribution. The ICC statistic can be applied within each subject, and the pooled index over K subjects can be computed as:

$$Z_{\mathrm{G}} = \left[\widehat{\mathrm{ICC}}_{\mathrm{G}} - \mathrm{ICC}_0\right] / \sqrt{\mathrm{Var}(\widehat{\mathrm{ICC}}_{\mathrm{G}})}\,, \tag{7}$$

where $\widehat{\mathrm{ICC}}_{\mathrm{G}} = \sum_j^K \omega_j \widehat{\mathrm{ICC}}_j$ with $\sum_j^K \omega_j = 1$; $\widehat{\mathrm{ICC}}_j$ denotes the estimate corresponding to the jth subject, and $\omega_j = (\frac{1}{Var(\widehat{\mathrm{ICC}}_j)})\mathrm{H}^{-1}$ with $\mathrm{H} = \sum_j^K \frac{1}{Var(\widehat{\mathrm{ICC}}_j)}$. The Z_{G} scores can be evaluated against a standard Gaussian or phase-randomization distribution with the FDR control over the false positive rate.

4. Empirical study

Empirical data were collected from an eyes-closed/open task with 30 healthy subjects (15 males, average age 22.50±3.46). The task requested subjects to close or open their eyes following an acoustical instruction with a total of 8-min duration (i.e., 4-min eyes closed followed by 4-min eyes open). On a 3T scanner, structural T1-weighted images were first acquired (TR/TE = 2530ms/3.30ms, flip angle = 7 deg, 192 slices, FOV 256 mm, phase 100%, and 1mm thick), followed by functional images acquired using echo-planar imaging (EPI) with a T2*-weighted gradient-echo sequence (TR/TE = 2000ms/30ms, flip angle 84 deg, 35 slices, FOV 192mm, and 3.4mm thick). Data pre-processing was carried out in SPM8, and T1-weighted high-resolution images were co-registered to the mean of realigned EPI images and spatially normalized with voxel size 2 x 2 x 3mm to the standard space as defined by the Montreal Neurological Institute (MNI) T1-weighted template. In the pre-processing stage, the polynomial trends and global mean of BOLD time courses were removed before data analysis. In the within-subject reliability assessment, the time course in each voxel was divided into 2 replicates (i.e., M = 2), each with a period of either eyes closed or eyes open, and the Z_{G} scores in (7) were computed for all voxels. Fig. 1 plots the empirical Z_{G} for all voxels along with the phase randomized Z_{G} distribution (10 replicates). As a comparison, the standard Gaussian distribution is also plotted in the figure. In the between-subject reliability assessment, the Z score in (6) was computed using time courses of the 30 subjects (M = 30). Fig. 2 plots the empirical Z scores for all voxels along with the phase randomized Z distribution (10 replicates). The within-subject phase randomized Z_{G} distribution is deviated from Gaussian, but the between-subject phase randomized Z distribution can be closely approximated by a Gaussian distribution. The two plots suggest that the within-subject analysis must rely on the phase-randomization distribution for thresholding voxels, but the between-subject analysis can use either phase-randomization or Gaussian distribution. Because subjects received the acoustic instruction of closing or opening their eyes, the supra-threshold voxels in the within-subject ICC maps indicate that BOLD responses in the primary auditory and motor cortices are more reliable between the eyes-closed and –open functional states. The default mode regions, motor and occipital-parietal cortices, and thalamus are more reliable between subjects. The cluster- and voxel-FDR suggest the same results in the between-subject analysis.

5. Conclusions

The Gaussian test based on cluster-FDR is computationally more efficient than the phase-randomization method and provides slightly more conservative test than the latter in the between-subject analysis. The proposed ICC statistic is particularly useful for studies on multi-site characterization of fMRI experimental paradigms (e.g., a few blocks of finger tapping or memory encoding tasks), and for studies involving long-term sensory stimulation, in which explanatory variables may not be easily accessible in a model-dependent method (e.g., free viewing of movies or listening to narrated stories/music under different conditions). We also found the statistic most informative about voxels activated in both experimental and control conditions out of those merely

activated in either of the two conditions in experiments involving more complicated cognitive tasks. The seed-voxels in studies on connectivity networks can be selected using those with peak-Z_G or Z scores in individual ROIs. Resting with eyes open and closed are two distinct functional states, a transition between which would supply information on both physiological and psychological characteristics of individual subjects. Since the last decade, several studies have addressed the relationship between eyes-closed and -open reactions, especially by comparing between the fMRI and electroencephalograph (EEG) results. For instance, BOLD responses in the visual cortex were shown to be negatively correlated with those in the thalamus. Further, EEG alpha power was positively correlated with BOLD responses in the thalamus, and negatively correlated with those in the visual cortex in either of the two conditions. According to the reliability assessment, BOLD responses in the most brain regions can exhibit heterogeneous patterns in the eyes-closed and open conditions. In the calcarine cortex, lingual gyrus, and superior occipital gyrus, for instance, eyes-closed and -open reactions are accompanied by respectively decreased and increased activities. The increased/decreased activity in different regions can be interpreted as switching from one functional state to the other. When subjects opened their eyes, some of the self-regulation processes, especially the interaction between cortical and subcortical structures, could have been actively suppressed. It is interesting to note that most brain regions exhibit heterogeneous reactions to eyes-closed and -open instructions (reliable regions in the between-subject analysis, but not in the within-subject analysis). The proposed ICC statistic satisfies the desirable feature of simplicity in fMRI applications without requiring a priori knowledge on task related BOLD responses. The potential use of the ICC statistic remains to be further investigated in more fMRI applications.

References

Bradley RC (2011). On the behavior of the covariance matrices in a multivariate central limit theorem under some mixing conditions. Ithaca, NY: Cornell University Library [Internet].

Dryden IL, Kume A, Le H, Wood ATW (2010). Statistical inference for functions of the covariance matrix in the stationary Gaussian time-orthogonal principal components model. Annals of the Institute of Statistical Mathemtics, 62:967–994.

Chumbley J, Worsley K, Flandin G, Friston K (2010). Topological FDR for neuroimaging. Neuroimage, 49:3057–3064.

Heller R, Stanley D, Yekutieli D, Rubin N, Benjamini Y (2006). Cluster-based analysis of fMRI data. NeuroImage, 33:599–608.

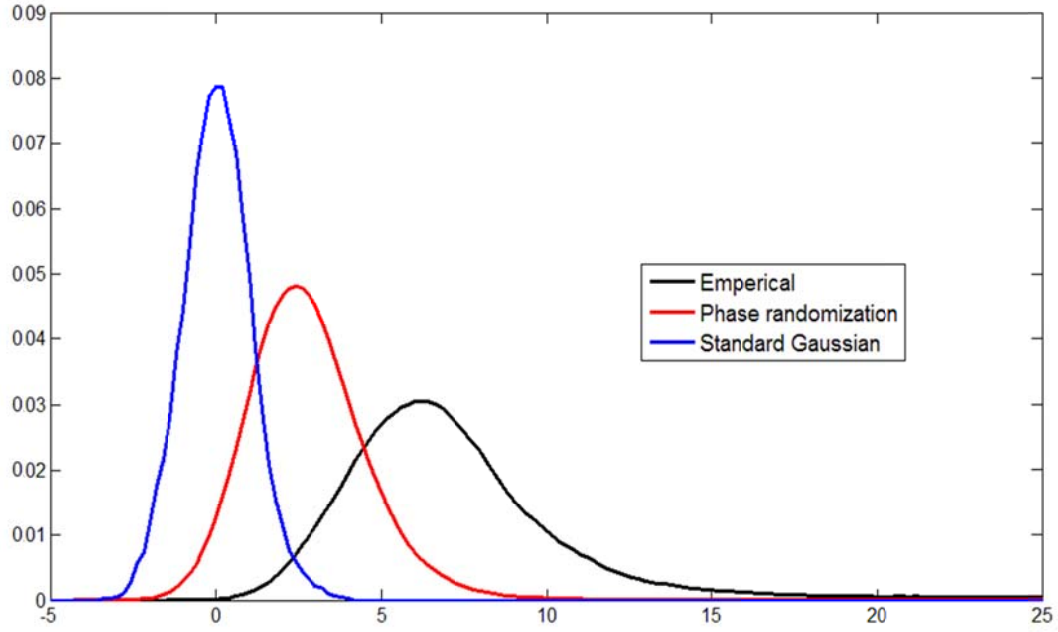

Fig. 1: Plots of empirical, phase-randomization and standard normal distributions in the within-subject analysis.

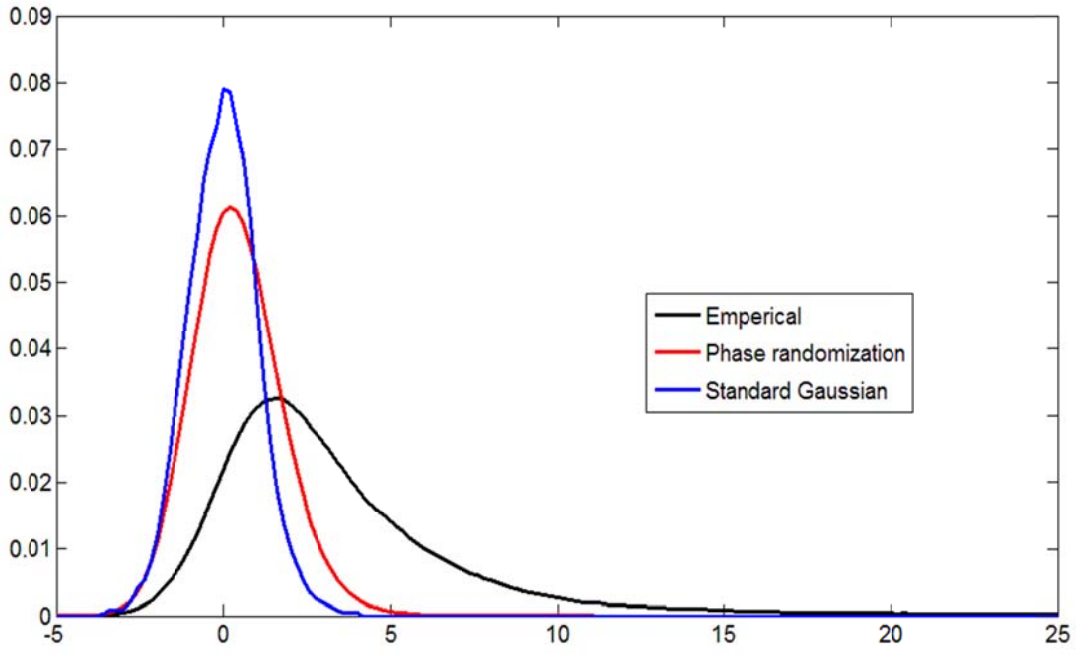

Fig. 2: Plots of empirical, phase-randomization and standard normal distributions in the between-subject analysis.