Statistica Sinica **18**(2008), 1253-1268

A BOOTSTRAP TEST TO INVESTIGATE CHANGES IN BRAIN CONNECTIVITY FOR FUNCTIONAL MRI

Pierre Bellec[1], Guillaume Marrelec[2] and Habib Benali[2]

[1]*McGill University and* [2]*Inserm U678/UPMC*

Abstract: Functional magnetic resonance imaging (fMRI) allows for the indirect measurement of whole brain neuronal activity using local blood oxygenation level. Functional connectivity, i.e., the correlation between the temporal activity of remote regions, may be used to track brain reorganization while, for example, a subject learns a new skill. However, testing the significance of changes in functional connectivity is challenging for individual data, because fMRI time series exhibit dependencies in both space and time that may not be properly captured by classical parametric models. To address this issue, we propose a new statistical procedure in a bootstrap hypothesis testing framework after various strategies were implemented to take temporal dependencies into account. These alternatives were evaluated on Gaussian and non-Gaussian Monte-Carlo simulations of space-time processes, as well as on a longitudinal study of motor skill learning. The results demonstrated that neglecting the temporal dependencies or modeling them as an autoregressive process of order 1 may lead to poor control of the false positive rate, i.e. to liberal tests. The version of the procedure based on a circular block bootstrap achieved robust, satisfactory performances in all settings.

Key words and phrases: Block bootstrap, correlation, data-driven block length selection, double bootstrap, fMRI, functional connectivity, hypothesis testing.

1. Introduction

Blood oxygen level dependent (BOLD) functional magnetic resonance imaging (fMRI) is a non-invasive technique that measures the hemodynamic correlates of whole-brain neural activity (Ogawa et al. (1990)). One main contribution of fMRI to brain science has been to help identify the brain regions engaged in the performance of a given task (Worsley and Friston (1995)). Beyond the mere localization of brain functions, fMRI has also contributed to the elucidation of some aspects of the interactions within a network of brain regions (Marrelec, Bellec and Benali (2006)). A popular approach in connectivity studies is the so-called functional connectivity, operationally defined as the correlation between the fMRI time series associated with two voxels or regions (Friston (1994)).

Assessing changes in functional connectivity has already proved to be a powerful tool to investigate brain reorganization (e.g., Dodel et al. (2005)). For this

purpose, statistical procedures have mostly been applied at a group level and therefore have dealt with independent measurements from different subjects, allowing use of general techniques such as the analysis of variance (ANOVA) (Morgan and Price (2004)). At the individual level, statistical tests have to deal with the uncertainty related to fMRI time series, which is more challenging. A first issue is that the assumption of a joint Gaussian distribution of individual fMRI time series is questionable (Gavrilescu et al. (2004)). A second issue is that, in functional connectivity studies, regions have a temporal activity dominated by the slow vascular response to neuronal activity. The associated fMRI time series thus cannot be regarded as independent and identically distributed (i.i.d.) samples from a given process (Bullmore et al. (2000)). While some resampling procedures have been put forward as remedies for these issues in the context of the general linear model (Friman and Westin (2005)), they are not readily applicable to functional connectivity studies.

In this paper, we propose a new statistical procedure designed to test changes in functional connectivity at the individual level. Various strategies were investigated to take temporal dependencies into account, and these alternatives were evaluated on Gaussian and non-Gaussian Monte-Carlo simulations of space-time processes. The procedure was also evaluated on an experiment of motor skill learning, with a longitudinal study of three subjects.

2. Bootstrap Hypothesis Testing Procedure

This section presents the details of the statistical testing procedure. The test is applied on a connectivity measure, i.e., a function of the fMRI data which can be, for example, the average functional connectivity measure (AFC), see Section 2.1. A bootstrap data-generating process (DGP) is used to approximate the distribution of the AFC, and multiple strategies can be implemented to take temporal dependencies into account, see Section 2.2. The bootstrap DGP is modified to conform with a null hypothesis in which the connectivity measures have the same distribution in two datasets. This approach, called bootstrap hypothesis testing, provides an estimate of the false positive rate (FPR) when rejecting the null hypothesis, see Section 2.3. Because multiple comparisons are performed, an estimate of the false discovery rate (FDR) is also implemented, see Section 2.4. Finally, a more complex version of the initial bootstrap algorithm is introduced to improve the accuracy of FPR and FDR estimation, see Section 2.5. The abbreviations used throughout the article have been listed in Supplementary Material, A.

2.1. Intra- and inter-networks average functional connectivity

Let $\mathbf{y}$ be a series of L functional datasets $(\mathbf{y}^l)_{l=1}^L$, assumed to have an identical size, $T \times N$, for simplicity. The time series $(y_{t,i}^l)_{t=1}^T$ is usually taken to

be the spatial average of the time series of all voxels within a region i and for an experimental condition l. A (multivariate) smooth function of $\mathbf{y}^l$ is derived to estimate a connectivity measure. For example, the functional connectivity between two regions i and j is the classic estimate of Pearson linear correlation coefficient $r_{ij}(\mathbf{y}^l)$. The AFC $\boldsymbol{\theta}$ is another measure that applies to a set of N brain regions grouped into M non-overlapping sets $(S_m)_{m=1}^M$, called networks. Many approaches can be used to define the regions and networks, e.g., a general linear model (Worsley and Friston (1995)), independent component analysis (McKeown et al. (1998)) or hierarchical clustering (Bellec et al. (2006)). The inter-networks AFC between S_m and $S_{m'}$ is defined as $(\#S_m \#S_{m'})^{-1} \sum_{i \in S_m}^{j \in S_{m'}} r_{ij}(\mathbf{y^l})$, where $\#$ is the cardinality of a set. The intra-network AFC limits the average to pairs of distinct regions: $2(\#S_m(\#S_m - 1))^{-1} \sum_{i,j \in S_m}^{i \neq j} r_{ij}(\mathbf{y^l})$. The (multivariate) AFC $\boldsymbol{\theta}(\mathbf{y})$ is the $M(M+1)/2$ vector of distinct intra- and inter-networks AFC measures placed in any arbitrary order, and $\theta(\mathbf{y})$ denotes any of these univariate measures.

2.2. Bootstrap data-generating processes

To test for significant changes in AFC between two fMRI datasets, the distribution of the random variable $\theta(\mathbf{y}^l)$ needs to be approximated. Formally, regional time series $\mathbf{y}^l$ are modelled as a sample from a N-dimensional stationary random process $\mathbf{Y}^l = (\mathbf{Y}_t^l; t \in \mathbb{Z})$ such that $(\mathbf{Y}_t^l)_{t=t_0+1}^{t_0+T}$ has a probability density function (pdf) f^l that depends on T but not on t_0, because of stationarity. Having observed $\mathbf{y}^l$, the bootstrap consists of building an approximation $\hat{f}_{\mathbf{y}}^l$ of f^l, which can be done under various assumptions. In practice, bootstrap estimates are built through Monte-Carlo sampling and it is not necessary to have an explicit expression for $\hat{f}_{\mathbf{y}}^l$, but rather to be able to draw samples $\mathbf{y}^{l,*}$ from $\hat{f}_{\mathbf{y}}^l$ through a DGP. In the classic independent and identically distributed (i.i.d.) case, the DGP consists of sampling independently T values $u(t)$ in $\{1, \ldots, T\}$ with a uniform probability $1/T$, with $y_{t,i}^{l,*}$ equals to $y_{u(t),i}^l$ for all $i = 1, \ldots, N$ (Efron and Tibshirani (1994)). This DGP, called **iidB**, respects the spatial dependence of the data, because the same temporal samples are used for all spatial locations, and it leads to consistent confidence intervals for the spatial correlations when the time series are i.i.d. (Shao and Tu (1995)).

Unfortunately, **iidB** is not suited for data with temporal dependencies, yet the DGP can be adapted to resample the data while preserving its intrinsic temporal structure. A class of DGPs, called pre-withening, is based on a parametric assumption on the temporal structure of $\mathbf{Y}$. The parameters of temporal dependencies are estimated to transform the data into a space where the i.i.d. assumption holds. For example, for a autoregressive process of order 1 (AR1), the DGP consists of estimating the parameter of the AR1 process and the i.i.d.

residuals, then applying **iidB** to the residuals. This semi-parametric DGP, called **AR1B**, has been advocated as an accurate fMRI data resampling scheme in the estimation of the null distribution in a general linear model (Friman and Westin (2005)).

A fully non-parametric alternative to **AR1B** is the circular block bootstrap (**CBB**) (Shao and Tu (1995)). The **CBB** consists of drawing blocks of the time series rather than independent observations in order to respect the temporal dependencies of the data. The block length h needs to be adapted to the range of temporal dependencies and the number of volumes T. For adequate h values, the **CBB** preserves spatial correlation, and formally leads to consistent confidence intervals of spatial correlations (Lahiri (2003)). In order to choose h, we propose the following maximum variance criterion (MVC), which indirectly intends to maximize the estimated FPR of the testing procedure by maximizing the variance of the bootstrap distribution. Specifically, the standard deviation $\hat{\sigma}_{\mathbf{y}^l}(h)$ of the bootstrap distribution of $\theta(\mathbf{y}^{l,*})$ is estimated by Monte-Carlo sampling on a grid of reasonable values for h; when multiple measures and datasets are considered, $\hat{\sigma}_{\mathbf{y}}(h)$ is the average of $\hat{\sigma}_{\mathbf{y}^l}(h)$ across all measures θ and datasets $\mathbf{y}^l$; the block length h_{opt} is selected as the one which maximises $\hat{\sigma}_{\mathbf{y}}(h)$.

2.3. Bootstrap estimate of the false positive rate

For a dataset $\mathbf{y} = (\mathbf{y}^l)_{l=1}^2$, i.e., $L = 2$, the statistic of interest, $\delta(\mathbf{y})$, is the difference in AFC between the two conditions, $\theta(\mathbf{y}^2) - \theta(\mathbf{y}^1)$. Under the null hypothesis $\mathcal{H}_0$, the pdfs f^1 and f^2 are identical, equal to f^0 say, which implies that the AFC measures have the same distributions. The cumulative distribution function (cdf) of $\delta(\mathbf{Y})$ under the null hypothesis is

$$G^0(x) = \Pr\left\{\delta(\mathbf{y}) \leq x | f^0\right\}, \quad \forall x \in \mathbb{R}. \tag{2.1}$$

In general, G^0 is monotone increasing (not strictly) and right-continuous, bounded between 0 and 1. Because δ is bounded, G^0 is moreover increasing from 0 to 1 on $[-2, 2]$. The left, one-tailed FPR $\alpha_{\mathbf{y}}$ under the null hypothesis $\mathcal{H}_0$ is $G^0\{\delta(\mathbf{y})\}$.

We propose to estimate $\alpha_{\mathbf{y}}$ by modifying the DGP in order to force the samples to conform with $\mathcal{H}_0$ even when the alternative hypothesis of different pdfs holds. This type of approach, called bootstrap hypothesis testing, has received surprisingly little attention compared to bootstrap confidence interval construction, but has still yielded compelling results (Martin (2007)). Under $\mathcal{H}_0$, the DGPs introduced in the last section can be modified to build samples of the estimated pdf $\hat{f}^0_{\mathbf{y}}$ in the following way. Because f^1 and f^2 are identical, bootstrap samples $\mathbf{y}^{1,*}$ and $\mathbf{y}^{2,*}$ can actually be drawn from a single distribution $\hat{f}^l_{\mathbf{y}}$. Half of the boostrap samples will be generated with $l = 1$ and the other half with $l = 2$.

The specifications of the null distribution apply to $\delta(\mathbf{y}^*) = \theta(\mathbf{y}^{2,*}) - \theta(\mathbf{y}^{1,*})$. Conversely, under such a DGP, the regional time series of the bootstrap data $\mathbf{y}^* = (\mathbf{y}^{l,*})_{l=1}^2$ have, by construction, an identical distribution, and thus conform to $\mathcal{H}_0$. The DGP under $\mathcal{H}_0$ can be used to build a bootstrap estimate of G^0:

$$\forall x \in \mathbb{R}, \quad \hat{G}^0_{\mathbf{y}}(x) = \Pr\left\{\delta(\mathbf{y}^*) \leq x | \hat{f}^0_{\mathbf{y}}\right\} \tag{2.2}$$

$$\doteq (B+1)^{-1} \# \left\{b = 1, \ldots, B | \delta(\mathbf{y}^{*b}) \leq x\right\}, \tag{2.3}$$

where # is the cardinality of a set. Equation (2.3) is a Monte-Carlo approximation of (2.2), and the symbol $\doteq$ means that the two terms are asymptotically equal as B tends toward infinity. Because of the discrete nature of the Monte-Carlo approximation, $\hat{G}^0_{\mathbf{y}}$ is a step discontinuous function. This has a practical disadvantage, because the algorithm we present in Section 2.5 requires the inversion of $\hat{G}^0_{\mathbf{y}}$. The application of a linear interpolation between discontinuity points on a grid of the possible values of $\delta(\mathbf{y})$, i.e. [-2,2], may be used to ensure continuity. In addition, imposing $\hat{G}^0_{\mathbf{y}}(-2) = 0$ and $\hat{G}^0_{\mathbf{y}}(2) = 1$) on the interpolation, together with the choice of a normalization by $(B+1)$ instead of the classical B in (2.3) which has asymptotically no impact, ensure that the new interpolated estimate, also denoted $\hat{G}^0_{\mathbf{y}}$ for simplicity, is monotone increasing (strictly on $[-2, 2]$) and continuous, and thus defines a bijection from $[-2, 2]$ on $[0, 1]$.

The bootstrap estimate $\hat{\alpha}_{\mathbf{y}}$ of the (left one-tailed) FPR under the null hypothesis is $\hat{G}^0_{\mathbf{y}}\{\delta(\mathbf{y})\}$, and the bootstrap estimate of the bilateral FPR is

$$\hat{p} = 2\,\min(\hat{\alpha}_{\mathbf{y}}, 1 - \hat{\alpha}_{\mathbf{y}}). \tag{2.4}$$

Note that the bootstrap scheme proposed here is not the only one that satisfies the null hypothesis requirements. The bootstrap samples $\mathbf{y}^{1,*}$ and $\mathbf{y}^{2,*}$ could, for example, be generated by mixing data samples from both $\mathbf{y}^1$ and $\mathbf{y}^2$ under a given DGP. The behavior of alternative schemes for the null hypothesis has not been investigated in this work.

2.4. Bootstrap estimate of the false discovery rate

The previous section focussed on a particular case where the connectivity measure was univariate and only two conditions were compared, leading to a single test. In general, there are $M(M+1)/2$ distinct AFC measures and an arbitrary number L of conditions $\mathbf{y} = (\mathbf{y}^l)_{l=1}^L$, implying $L(L-1)/2$ possibly non-redundant comparisons. This raises the issue of multiple comparisons, precisely $M(M+1)L(L-1)/4$ of them; moreveover these are dependent because the same data is involved in more than one univariate AFC measure and more than one comparison. The FDR initially proposed by Benjamini and Hochberg (1995) is

an approach to this issue that is very well suited to bootstrap hypothesis testing. The key idea is to allow for some false positives to arise in the procedure, but to relate their number to the total number of positive findings. For a given FPR threshold p on each individual test, let $D_{\mathbf{y}}(p)$ be the total number of discoveries, i.e., the number of tests on $\theta(\mathbf{y}^l) - \theta(\mathbf{y}^{l'})$ where $\hat{p}$ is smaller than p. Among these $D_{\mathbf{y}}(p)$ discoveries, there are $D^F_{\mathbf{y}}(p)$ unknown false positive discoveries and $D^T_{\mathbf{y}}(p)$ true positive discoveries. The FDR $q(p)$ is defined as the expectation $\mathbb{E}\left\{D^F_{\mathbf{y}}(p)/D_{\mathbf{y}}(p)|f\right\}$, where f is the joint pdf of $\mathbf{Y} = (\mathbf{Y}^l)_{l=1}^L$. Let $\mathcal{G}_0$ be the global null hypothesis that all data $(\mathbf{Y}^l)_{l=1}^L$ have the same distributions f^0. Samples $(\mathbf{y}^{l,*})_{l=1}^L$ from $\hat{f}^0_{\mathbf{y}}$ under the global null can be generated from a single dataset $\mathbf{y}^l$, the value l taking random values in $\{1,\dots,L\}$ with uniform probability for each bootstrap sample. Let $D^*_0(p)$ be the number of false positives of one sample $\mathbf{y}^*$ under the global null hypothesis $\mathcal{G}^0$. Logan and Rowe (2003) proposed the following conservative estimate of the FDR:

$$\hat{q}(p) = \mathbb{E}\left\{\frac{D^*_0(p)}{D^*_0(p) + \hat{D}_T(p)}|\hat{f}^0_{\mathbf{y}}\right\} \doteq B^{-1}\sum_{b=1}^{B}\left\{\frac{D^{*b}_0(p)}{D^{*b}_0(p) + \hat{D}_T(p)}\right\}, \tag{2.5}$$

where $\widehat{D}_T(p) = D(p) - pM(M+1)L(L-1)/4$. To achieve a comparison at a given FDR threshold, e.g., $q < 0.05$, the largest FPR threshold p such that $\hat{q}(p) < 0.05$ is selected.

2.5. Yet another double bootstrap algorithm

Despite its asymptotic consistency, the simple bootstrap estimate $\hat{p}$ of the FPR introduced in Section 2.3 may be too liberal on finite time series, which would also compromise the estimate $\hat{q}(p)$ of the FDR. Some procedures have been proposed to correct for the fact that all bootstrap samples are actually generated from the same dataset. Such correction usually results in a better behavior on finite samples, and a faster asymptotic convergence. Instances of correction procedures are the double bootstrap (**DB**) (Hall and Martin (1998)), computationally demanding, and the fast double bootstrap (**FDB**) (Davidson and Mackinnon (2007)) which, as the name suggests, is faster than **DB**. We present here a new algorithm called "yet another double bootstrap" (**YADB**) whose complexity is in general of the same order as the **DB**, but can be made markedly faster than the **FDB** in the particular context of bootstrap hypothesis testing on a difference.

The **YADB** appears as an a posteriori correction of the simple bootstrap estimate of FPR. The theoretical motivation behind this correction is that it would be desirable for the bootstrap cdf estimates to be distributed equally around the true cdf, i.e.,

$$\forall x \in [-2,2], \quad H(x) = \mathrm{med}(\hat{G}^0_{\mathbf{y}}(x)|f^0) = G^0(x), \tag{2.6}$$

where med denotes median. Note that each cdf $\hat{G}^0_{\mathbf{y}}$ is a continuous monotone bijection from $[-2, 2]$ on $[0, 1]$, and H thus statisfies the same properties. This allows us to define the function ξ as $G^0 \circ H^{-1}$ which satisfies $\xi \circ H(x) = G^0(x)$, and ξ is a continuous monotone increasing bijection from [0,1] to [0,1]. The corrected estimate of cdf defined as $\xi \circ \hat{G}^0_{\mathrm{y}}$ satisfies the condition (2.6), because the median and a monotone increasing function commute. Unfortunately, the function ξ is unknown, yet it can be estimated through bootstraping. Informally, the idea is to compare the (simple) bootstrap cdf to the median cdf derived when a bootstrap sample is used instead of the original dataset. The function which brings the median double bootstrap cdf on the simple bootstrap cdf will serve as an estimate of the one which brings the single bootstrap cdf on the true cdf. Formally, the median double boostrap cdf is

$$\hat{H}_{\mathbf{y}}(x) = \mathrm{med}(\hat{G}^0_{\mathbf{y}^*}(x)|\hat{f}^0_{\mathbf{y}}) \doteq \mathrm{med}\left\{\hat{G}^0_{\mathbf{y}^{*\mathbf{c}}}(x),\, c = 1, \ldots, C\right\}. \tag{2.7}$$

The estimate $\hat{\xi}$ of ξ is then $\hat{G}^0_{\mathbf{y}} \circ \hat{H}^{-1}_{\mathbf{y}}$, with numerical inversion and composition performed through linear interpolation on a grid of $[0, 1]$. The corrected cdf $\hat{\xi} \circ \hat{G}^0_{\mathrm{y}}$ can be used to derive some **YADB** estimates of the FPR and FDR in the same way as it was done with the simple bootstrap.

Regarding computational complexity, the **DB** requires approximately BC samples $\mathbf{y}^*$ while the **FDB** involves $2B$ samples $\mathbf{y}^*$, with B at least $1,000$ and C at least 50 (Hall, Lee and Young (2000) and Davidson and Mackinnon (2007)). The **YADB** algorithm requires one initial cdf estimation $\hat{G}^0_{\mathbf{y}^*}$ followed by the generation of C samples $\mathbf{y}^*$ and the corresponding cdf estimation $\hat{G}^0_{\mathbf{y}^*}$, see (2.7), each cdf estimation requiring B samples $\mathbf{y}^*$, see (2.3). This approach thus involves $B + C(B + 1)$ Monte-Carlo samples $\mathbf{y}^*$, with $B > 1,000$ and C at least 20 in order to stabilize the estimation of the median. This complexity is asymptotically equivalent to that of the **DB**.

Taking advantage of the fact that the statistic δ is a difference, a simple computational trick can be used to drastically cut down on the computational cost of the FPR estimation for one comparison, e.g., $\theta(\mathbf{y}^2) - \theta(\mathbf{y}^1)$. Let $(\mathbf{y}^{*d})_{d=1}^{D}$ be some independent samples generated under $\mathcal{H}_0$ from a single dataset, say $\mathbf{y}^1$. The differences $\{\theta(\mathbf{y}^{2,*d}) - \theta(\mathbf{y}^{1,*d'})\}$ are distinct samples of the type $\delta(\mathbf{y}^*)$ for all $D(D-1)$ pairs where $d \neq d'$. The samples $\delta(\mathbf{y}^*)$ are not independent but this has small impact in practice. For deriving a cdf with B samples $\delta(\mathbf{y}^*)$, $B/2$ samples are generated from $\mathbf{y}^1$ and another $B/2$ from $\mathbf{y}^2$, each batch requiring about $D = 1/2 + \sqrt{1 + B/2}$ effective samples $\mathbf{y}^*$ (more precisely the upper integer part). Using this trick to compute the cdfs, the number of samples $\mathbf{y}^*$ can be reduced to approximately $B\sqrt{2C}$ with the **DB** and $C\sqrt{2B}$ with the **YADB**, but the same idea does not apply to the **FDB** due to implementation details, leaving

its complexity at $2B$. In practice, for a reasonably accurate estimation of a 0.05 FPR with $B = 10,000$ and $C = 50$, the **DB** would use approximately $100,000$ bootstrap samples $\mathbf{y}^*$, the **FDB** would use $20,000$ and the **YADB** $7,072$.

3. Simulation Study

3.1. Description of the experiments

The behavior of the testing procedure was investigated on Monte-Carlo simulations using different alternatives for the bootstrap DGP, **iidB**, **AR1B** and **CBB**-and using two models of multivariate space-time processes-a Gaussian process separable in space and time (**GSST**) and a non-Gaussian hidden Markov multi-states (**HMMS**) process. The analytic expression and validity of models of time correlation (tC) and space correlation (sC) are reported in Supplementary Material, Section B, and only the values of parameters are listed here. The parametric tC model was exponential, equivalent to an AR1 model, with parameter $a = 0.5$. The sC model was homogeneous, which means that the sC associated with regions of two networks m and m' was constant, equal to an AFC parameter $\theta_{m,m'}$. We used $M = 3$ networks, each one formed of 5 regions to limit computation time. Three scenarios were considered for the sC, with most AFC measures identical in all scenarios: $\theta_{11} = \theta_{22} = \theta_{33} = 0.6$, $\theta_{12} = \theta_{13} = 0.15$, and the AFC between networks 2 and 3 varying, $\theta_{23} \in \{-0.15, 0, 0.15\}$ for each scenario. The **HMMS** model in addition implemented the idea that two different systems could perform the same task, and that the brain may sometimes switch from one to the other. Formally, the sC at each time frame in the **HMMS** model actually took one of two possible values with θ_{12} and θ_{13} either equal to -0.05 or 0.35, depending on the state of a binary hidden Markov chain. The probability that the state of the chain changed from one time frame to the next was 0.05. For each simulation, three datasets $(\mathbf{y}^{\mathbf{l}})_{l=1}^{3}$ of size $T \times N$ were generated, each corresponding to one of the three scenarios of tC and sC.

A first batch of simulations was used to investigate how the choice of the optimal block length h_{opt} was related to the number of time samples T (50, 100 or 200) and the simulation model (**GSST** or **HMMS**). Simulated datasets $(\mathbf{y}^{\mathbf{l}})_{l=1}^{3}$ were used to select the block length in **CBB** using the MVC, with $B = 300$ bootstrap samples, and h in $\{1, 4, 7, 10, 20, 30, 40, 50, 75, 100\}$. There were 100 simulations performed for each T and simulation type, for a total of 600 simulations.

A second batch of simulations was used to investigate the sensitivity and specificity of the procedure. Specifically, simple bootstrap and the **YADB** algorithm were applied to test for changes in AFC between $\mathbf{y}^2$ and $\mathbf{y}^1$ (hard comparison, $\theta_{23}^2 - \theta_{23}^1 = 0.15$), and between $\mathbf{y}^3$ and $\mathbf{y}^1$ (easy comparison, $\theta_{23}^3 - \theta_{23}^1 = 0.3$),

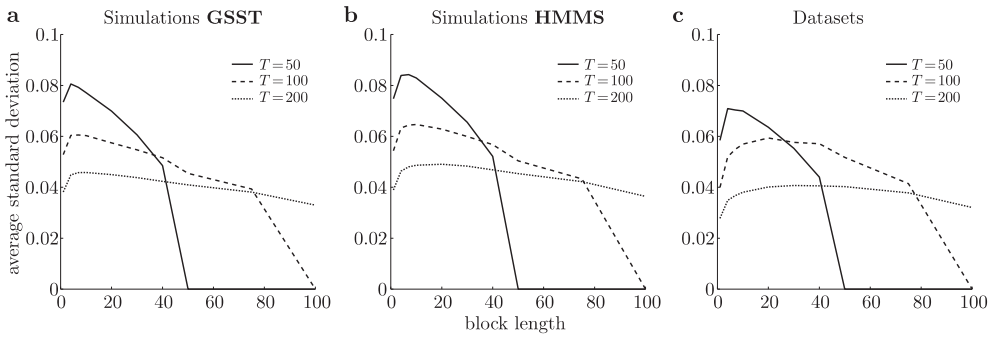

Figure 3.1. Average bootstrap estimate of the average standard deviation of AFC measures $\hat{\sigma}_{\mathbf{y}}(h)$ for different numbers of time samples T. Curves have been averaged across all simulations for a given simulation type, and all subjects for real data. Note the smooth relationship between h and $\hat{\sigma}_{\mathbf{y}}(h)$, achieving a single global maximum for a value h_{opt} which is dependent on T and the data type.

with $B = 10,000$ bootstrap samples for the simple bootstrap and the FDR estimation, and $C = 25$ iterations with $B = 5,000$ samples each in the **YADB** bootstrap. For each number of time frames ($T \in \{50, 100, 200\}$), each DGP (**CBB**, **AR1B**, **iidB**) and each simulation type (**GSST**, **HMMS**), 500 Monte-Carlo simulations were done, for a total of $9,000$ simulations. For every possible configuration, the effective FPR $\hat{e}$ was estimated by deriving the detection rate of differences in θ_{12} and θ_{13} with an estimated FPR $\hat{p} < 0.05$. Confidence interval at the 90% level on $\hat{e}$, symmetric and bilateral, were derived using the asymptotic approximation of the variance $\hat{e}(1-\hat{e})/D$, where D is the number of Monte-Carlo samples: 500 (samples) $\times$ 2 (measures) $\times$ 2 (comparisons). The effective FDR for all AFC measures was assessed by deriving the average ratio between the number of false positives and the total number of discoveries over all simulations when an FDR threshold $\hat{q} < 0.05$ was applied. Confidence intervals at the 90% level on the effective FDR were derived using simple bootstrap of the Monte-Carlo samples. Finally, by considering a grid of thresholds for the FPR covering $[0, 1]$, a receiver-operating characteristic (ROC) curve of the effective sensitivity as a function of the effective specificity was derived for each DGP, each type of simulations, and each type of comparison (easy or hard), with $T = 200$.

3.2. Results

Figure 3.1(**a**,**b**) shows the bootstrap estimate of the average standard deviation of AFC measures $\hat{\sigma}_{\mathbf{y}}(h)$ as a function of the block length h, averaged across all simulations $\mathbf{y}$. The standard deviation smoothly increased then decreased,

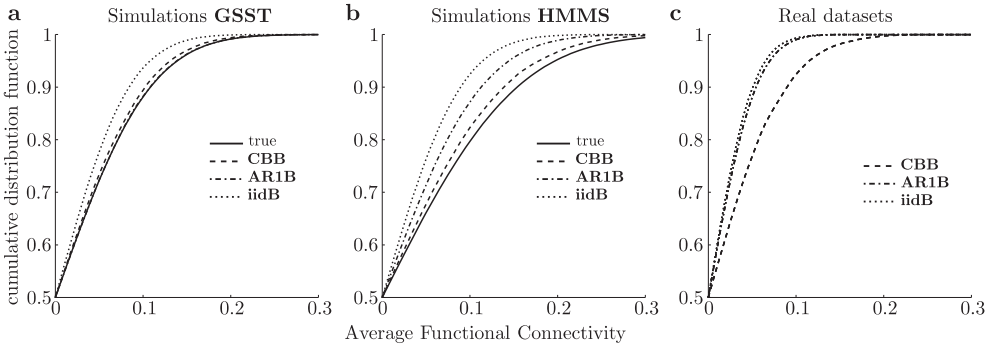

Figure 3.2. Median cumulative distribution function (cdf) under the null hypothesis estimated through **YADB**, with $T = 200$. Note that for **GSST** simulations, **AR1B** and **CBB** achieved more accurate approximations than **iidB**, the **AR1B** cdf actually overlapping the true cdf. With **HMMS** simulations, both **AR1B** and **iidB** underestimated the variance of the distributions, while **CBB** achieved a reasonable approximation. Results on data suggested that **AR1B** and **iidB** would have a similar behavior and both underestimated the variance of the distribution, compared to **CBB**.

admitting one single global maximum for $h = h_{\text{opt}}$. The value h_{opt} increased with increasing length of time series, and was larger for **HMMS** than **GSST** simulations: for **GSST** simulations with $T \in \{50, 100, 200\}$, the distribution of h_{opt} exhibited a median of $4, 7, 10$, a 0.05-unilateral lower percentile of $4, 4, 4$ and a 0.05-unilateral upper percentile of $10, 20, 30$; for **HMMS** simulations with $T \in \{50, 100, 200\}$, the distribution of h_{opt} exhibited a median of $7, 10, 20$, a 0.05-unilateral lower percentile of $4, 4, 7$ and a 0.05-unilateral upper percentile of $10, 20, 40$. When a large number of time samples was simulated ($T = 200$), the standard deviation was found stable on a large interval of values h around h_{opt}.

Figure 3.2(**a**,**b**) presents the median across all simulations and comparisons of the **YADB** estimate of the cdf under the null hypothesis for θ_{12} and θ_{13}. On **GSST** simulations, the **AR1B** DGP produced a very accurate estimation of the true median cdf. The **CBB** also generated a satisfactory approximation, while **iidB** performed badly. On **HMMS** simulations, none of the DGP produced very accurate approximations of the true median cdf, yet the DGP **CBB** performed best and **iidB** performed worst. Both **AR1B** and **iidB** approximations were drastically underestimating the actual variance of the distribution, suggesting liberal statistical tests. These results were confirmed by examining the effective FPR of the testing procedures.

With $T = 200$, $\hat{p} < 0.05$, and simulations **GSST**, the effective FPR $\hat{e}$ was 0.077 (90%-confidence interval $[0.066, 0.089]$) with **CBB**, 0.051 ($[0.041, 0.06]$)

with **AR1B**, and 0.133 ([0.118, 0.148]) with **iidB**. Both **CBB** and **AR1B** thus reached an acceptable efffective FPR, meaning $\hat{e} < 0.1$. On **HMMS** simulations, $\hat{e}$ was 0.098 ([0.085, 0.111]) with **CBB**, 0.161 ([0.145, 0.177]) with **AR1B**, and 0.256 ([0.236, 0.276]) with **iidB**. On this last type of simulations, **CBB** was thus the only DGP offering a reasonable control of the FPR. More comprehensive results can be found in Supplementary Material, Figure C.1. The results on effective FDR were very similar to the ones just stated above on FPR, see Figure C.2 of Supplementary Material. Despite the good control of false positives, the **CBB** was the DGP which performed the worst on ROC curves. However, all three DGP had very close performance, with a sensitivity of about $30 \pm 10\%$ for hard comparisons, and a sensitivity of $80 \pm 10\%$ for easy comparisons when an effective FPR (specificity) of 0.05 was considered, regardless of the simulation type, see Figure C.3 of Supplementary Material.

4. Application to fMRI Data: Motor Skill Learning

4.1 Description of the experiment

The **YADB** algorithm was evaluated on an experiment of motor skill learning, approved by the local ethic committee, with three right-handed, healthy male volunteers (age 25 to 27). Functional data was acquired while subjects steadily performed some motor sequences at a fixed, comfortable rate of 2 Hz. Three sequences were performed after one month of daily training (*known* conditions), and one at the very early stage of learning (*new* condition). Our first evaluation hypothesis was that similar networks would be engaged in the *known* conditions while marked differences in networks would be observed when comparing the *new* and the *known* conditions. As a positive control, we also compared these steady-state conditions to a block-designed condition alternating motor sequences and rest. The evaluation hypothesis regarding these comparisons was that, because of stimulus-locked fluctuations, the *block* condition should systematically exhibit larger correlations than those observed in the steady-state conditions within motor-related regions.

Functional data were acquired on a Bruker 3.0T MRI scanner at the fMRI Center in Marseille, France, with the following parameters. For each condition, 200 full brain volumes were recorded using a single-shot echo-planar imaging sequence (TR/TE = 2,333/30 ms, 64 × 64 matrix, 42 contiguous slices, FOV = 192 mm × 192 mm, slice thickness 3 mm, and flip angle = 81°). A high-resolution T_1-weighted scan was also acquired using the following MPRAGE sequence: TR = 11.6 ms, TE = 5.67 ms, TI = 800 ms, 256 × 192 × 104 matrix, FOV = 256 mm × 230 mm × 182 mm, and flip angle = 30°.

4.2. fMRI data analysis

The functional data were first corrected for delay in slice timing and inter-run motion using SPM2 (Friston and Worsley (1995)). The time series were also corrected for slow time drifts by application of a Butterworth high-pass filter with a cut-off frequency of 0.01 Hz (Smith et al. (1999)). Intra-run motion and physiological noise were reduced using a procedure based on independent component analysis (Perlbarg et al. (2007)). Regions were identified separately for each subject using the large-scale network identification (LSNI) method applied to a block dataset different from the one used in the following analysis (Bellec et al. (2006)). Three networks were reproducibly found across subjects, and the thirty regions with highest average functional connectivity were selected for each network: the first was composed of regions in large part located in the motor system that overlapped the regions expected to be activated by the task; the second included frontal cortex and bilateral superior parietal cortex; the third overlapped with the so-called default mode network as reported by Greicius et al. (2003).

We performed 10 statistical comparisons on the 6 AFC measures (3 intra-networks AFC and 3 inter-networks AFC), for a total of 60 tests per subject. First, we compared datasets acquired in the *known* conditions (3 *known-known* comparisons), then we compared the dataset in the *new* condition to the ones acquired in the *known* conditions (3 *new-known* comparisons), and finally, we compared the dataset acquired in the *block* condition to the ones acquired in the steady-state conditions (4 *block-steady* comparisons). For each subject, the optimal block length h_{opt} was selected using $B = 1,000$ bootstrap samples and the MVC criterion for $h \in \{1, 4, 7, 10, 20, 30, 40, 50, 75, 100\}$. We applied the **YADB** algorithm with each DGP, i.e. **CBB**, **AR1B** and **iidB**, $B = 100,000$ bootstrap samples for the level-1 bootstrap and FDR estimation, and $C = 100$ iterations of the bootstrap with $10,000$ samples each. The **CBB** DGP was applied three times with h in $\{30, 40, 50\}$ in order to cover the different values h_{opt} estimated on data. A test was considered significant for an estimated FPR $\hat{p} < 0.05$, and the associated FDR was derived for each subject.

4.3 Results

Figure 3.1**c** shows the bootstrap estimate of the average standard deviation of AFC measures $\hat{\sigma}_{\mathbf{y}}(h)$ as a function of the block length h, averaged across all subjects, and with the time series truncated to achieve $T = 50$, $T = 100$ and $T = 200$. The relationship between $\hat{\sigma}_{\mathbf{y}}(h)$ and h was smooth, and very similar in shape to the one observed on simulated data, see Figure 3.1**a,b**. However, the

values $\hat{\sigma}_{\mathbf{y}}(h)$ were smaller with actual data than with simulations, which may be due to the fact that simulations included too much temporal autocorrelation. The values estimated for h_{opt} were slightly higher than the ones observed on **HMMS** simulations, and consistent across subjects: 4, 4, 7 ($T = 50$); 20, 20, 20 ($T = 100$); and 30, 50, 30 ($T = 200$), for subjects 1 to 3, respectively.

Figure 3.2**c** presents the median cdf under the null hypothesis estimated through the testing procedure for all subjects and AFC measures, each curve corresponding to a choice of GDP. Compared to the cdf derived on simulations, see Figure 3.2**a,b**, the cdf on the datasets corresponded to distributions with smaller variance, which confirmed the previous results regarding block length selection. The median cdf derived using the DGP **AR1B** and **iidB** were very close, and departed markedly from the one derived using the **CBB**, which corresponded to a distribution with larger variance. This result suggested that the **CBB** would lead to much more conservative statistical tests than the ones performed using the **AR1B** or **iidB**, and that the difference in specificity would be even more drastic than it was on **HMMS** simulations. This result was confirmed by further examination of significant differences derived for each DGP.

Summaries of the tests performed with a FPR threshold of 0.05 are listed below for subjects 1 to 3, respectively. Note that the expected number of false positives per subject would be three if tests were independent and the estimates of FPR were exact. The **CBB** DGP with $h = 40$ led to an estimated FDR of 0.15, 0.14 and 0.29. Only a small number of significant differences were found in *known-known* comparisons (0, 3, 0). By contrast, a higher number was observed in *new-known* comparisons (1, 4, 2), and a large number was observed in *block-steady* comparisons (13, 8, 5). Interestingly, the AFC within the motor-related regions was found higher in the *block* condition than in any steady-state condition for the three subjects. These results, derived with the **CBB**, supported our evaluation hypothesis, yet the sensitivity in *new-known* comparisons appeared limited. The **AR1B** DGP led to smaller estimate of FDR than the **CBB** (0.09, 0.1, 0.16). Unfortunately, a large number of significant differences per subject were found in all comparison types: 15, 18, 39 total significant differences for *known-known*, *new-known* and *block-steady* comparisons, respectively, which did not support the evaluation hypothesis. The **iidB** and **AR1B** led to very similar results: the FDR values were 0.09, 0.09, 0.15 and the total numbers of significant differences per subject were 16, 20, 40 for *known-known*, *new-known* and *block-steady* comparisons, respectively, which again did not support our evaluation hypothesis.

To assert the robustness of the **CBB** results with respect to the block length h, we computed the absolute difference between the FPR estimated for $h = 40$

and $h \in \{30, 50\}$. The median of those values was 0.0153, 90% of the differences being smaller than 0.06, and the largest difference being 0.15. The number of significant differences, added for all subjects, was 3, 8, 27; 3, 7, 26; 2, 5, 25 for *known-known*, *new-known* and *block-steady* comparisons, respectively, and for h equal to 30, 40 and 50, respectively. The selection of a block length in this range thus did not qualitatively change the conclusions stated above, yet longer block length apparently led to slightly more conservative tests.

5. Conclusions

In this paper, we proposed a bootstrap hypothesis testing procedure designed to assess significant changes in functional connectivity between two fMRI datasets. A number of alternative bootstrap data-generating processes could be used within **YADB**, which differed by their approach of temporal dependencies: one neglected them (**iidB**), one assumed an autoregressive model of order 1 (**AR1B**), and the last one addressed them through block resampling (**CBB**).

The alternative data-generating processes were evaluated on Monte-Carlo simulations with space and time dependencies, and two simulation types were considered: a Gaussian model separable in space and time, and a non-Gaussian model exhibiting nonlinear temporal dependencies. The **iidB** scheme did not allow for a satisfactory control of false positive rate in any simulation type. The **AR1B** scheme performed very well on Gaussian simulations, but lead to very liberal tests on non-Gaussian simulations. The **CBB** scheme had a robust, satisfactory behavior in all settings, as soon as the number of time samples reached 200.

The testing procedure was also applied on an experiment of motor skill learning. The experimental design was such that we had a strong evaluation hypothesis regarding the existence or absence of significant differences. The **AR1B** and **iidB** schemes had a very similar behavior, which did not occur with simulations, and turned out to be far too liberal based on our evaluation hypothesis. The tests based on the **CBB** scheme were drastically more conservative, and the results conformed well with the evaluation hypothesis.

Together, our experiments on simulations and data suggest that fMRI individual data significantly depart from a joint Gaussian model separable in time and space with an autoregressive temporal structure. A fully non-parametric procedure taking temporal dependencies into accout is thus desirable to derive correct tests on functional connectivity at the individual level. Our results demonstrate that a bootstrap hypothesis testing approach using the **YADB** algorithm and a **CBB** bootstrap data-generating process is an adequate solution to this problem. An implementation of this procedure will be made available on the internet at http://wiki.bic.mni.mcgill.ca/index.php/BootstrapHypothesisTesting.

Acknowledgements

The authors are grateful to three anonymous referees for their insightful suggestions, to Jean-Luc Anton for the fMRI motor learning dataset, to Julien Doyon for helping in the design of the real experiment, and to Mélanie Pélégrini-Issac and Jonathan Lau for editing earlier versions of the manuscript.

References

Bellec, P., Perlbarg, V., Jbabdi, S., Pélégrini-Issac, M., Anton, J., Doyon, J. and Benali, H. (2006). Identification of large-scale networks in the brain using fMRI. *NeuroImage* **29**, 1231-1243.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57**, 289-300.

Bullmore, E., Horwitz, B., Honey, G., Brammer, M., Williams, S. and Sharma, T. (2000). How good is good enough in path analysis of fMRI data? *NeuroImage* **11**, 289-301.

Davidson, R. and Mackinnon, J. G. (2007). Improving the reliability of bootstrap tests with the fast double bootstrap. *Comput. Statist. Data Anal.* 51, 3259-3281.

Dodel, S., Golestani, N., Pallier, C., Elkouby, V., Le Bihan, D. and Poline, J.-B. (2005). Condition-dependent functional connectivity: syntax networks in bilinguals. *Philosophical Transactions of the Royal Society of London B* **360**, 921-935.

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York.

Friman, O. and Westin, C. F. (2005). Resampling fMRI time series. *NeuroImage* **25**, 859-867.

Friston, K. J. (1994). Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping* **2**, 56-78.

Friston, K. J. and Worsley, K. J. (1995). Analysis of fMRI times series revisited – again. *Neuroimage* **3**, 173-181.

Gavrilescu, M., Stuart, G. W., Waites, A., Jackson, G., Svalbe, I. D. and Egan, G. F. (2004). Changes in effective connectivity models in the presence of task-correlated motion: an fMRI study. *Human Brain Mapping* **21**, 49-63.

Greicius, M. D., Krasnow, B., Reiss, A. L. and Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proc. Natl. Acad. Sci. USA* **100**, 253-258.

Hall, P. and Martin, M. A. (1998). On bootstrap resampling and iteration. *Biometrika* **75**, 661-671.

Hall, P., Lee, S.M.S. and Young, G.A. (2000). Importance of interpolation when constructing double-bootstrap confidence intervals. *J. Roy. Statist. Soc. Ser. B* **62**, 479-491.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*. Springer, New York.

Logan, B. R. and Rowe, D. B. (2003). An evaluation of thresholding techniques in fMRI analysis. *Neuroimage 22*, 95-108.

Marrelec, G., Bellec, P. and Benali, H. (2006). Exploring large-scale brain networks in functional MRI. *J. Physiology, Paris* **100**, 171-181.

Martin, M. A. (2007). Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties. *Comput. Statist. Data Anal.* **51**, 6321-6342.

McKeown, M. J., Makeig, S., Brown, G. G., Jung, T. P., Kindermann, S. S., Bell, A. J. and Sejnowski, T. J (1998). Analysis of fMRI data by blind separation into independent spatial components. *Human Brain Mapping* **6**, 160-188.

Morgan, V. L. and Price, R. R. (2004). The effect of sensorimotor activation on functional connectivity mapping with MRI. *Magnetic Resonance Imaging* **22**, 1069-1075.

Ogawa, S., Lee, T. M., Kay, A. R. and Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc. Natl. Acad. Sci. USA* **87**, 9868-9872.

Perlbarg, V., Bellec, P., Anton, J.-L., Pelegrini-Issac, M., Doyon, J. and Benali, H. (2007). CORSICA: correction of structured noise in fMRI by automatic identification of ICA components. *Magnetic Resonance Imaging* **25**, 35-46.

Shao, J. and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Smith, A. M., Lewis, B. K., Ruttimann, U. E., Ye, F. Q., Sinnwell, T. M., Yang, Y., Duyn, J. H. and Frank, J. A. (1999). Investigation of low frequency drift in fMRI signal. *NeuroImage* **9**, 526-533.

Worsley, K. J. and Friston, K. J. (1995). Analysis of fMRI time-series revisited—again. *NeuroImage* **2**, 173-181.

McConnell Brain Imaging Center, Montreal Neurological Institute, McGill University, Montréal, H3A 2B4, Canada.

E-mail: pbellec@bic.mni.mcgill.ca

Inserm, UMR S 678, Laboratoire d'Imagerie Fonctionnelle, F-75634, Paris, France.

UPMC Univ Paris 06, UMR S 678, Laboratoire d'Imagerie Fonctionnelle, F-75634, Paris, France.

E-mail: guillaume.marrelec@imed.jussieu.fr

Inserm, UMR S 678, Laboratoire d'Imagerie Fonctionnelle, F-75634, Paris, France.

UPMC Univ Paris 06, UMR S 678, Laboratoire d'Imagerie Fonctionnelle, F-75634, Paris, France.

E-mail: habib.benali@imed.jussieu.fr

(Received April 2007; accepted February 2008)