Title: Hypothesis testing in the presence of noisy experimental replications

Authors: Diego Vidaurre[1*], Mark W. Woolrich[1], Theodoros Karapanagiotidis[2], Jonathan Smallwood[2], Thomas E. Nichols[3]

[1] Wellcome Trust Centre for Integrative Neuroimaging, Oxford Centre for Human Brain Activity, University of Oxford, UK.
[2] Department of Psychology, University of York, UK.
[3] Big Data Institute, University of Oxford, UK
* email: diego.vidaurre@ohba.ox.ac.uk

Abstract: We propose a simple procedure based on permutation testing that provides a way of combining the results from many individual tests that refer to the same hypothesis. This is needed when testing a measure whose value is obtained from a noisy process, which can be repeated multiple times, referred to as replications. Examples of a noisy process can be: (i) computational, e.g. when using an approximate inference algorithm (e.g. ICA) for which different runs can produce different results or (ii) observational, if we have the capacity to acquire data multiple times, and the different acquired data sets can be considered noisy examples of the underlying data that we are attempting to estimate; that is, we are not interested in the individual replications but on the unobserved process behind. This method can also be used when we intend to test multiple hypotheses, each with access to various replications, while correcting for the familywise error rate. Using both simulations and real data, we show that the proposed approach compares favourably to more standard approaches to this problem.

Introduction

Let us suppose that we are interested in testing hypotheses about variables, or set of variables, that we can observe on multiple occasions such that we end up having a number of noisy measures of the same underlying (unobserved) feature or process. This can happen when we replicate a measurement on multiple occasions for each subject, or if the design of the experiment is such that the repetitions are independent from each other (which would not be the case, for example, if there is a strong effect of learning or habituation across runs). This can also happen when we are modelling data using an approach that is complex enough that inferences about the model parameters can be slightly different every time we estimate the model, e.g. with different arbitrary initialisations. This is the case, for example, for independent component analysis (ICA, Hyvärinen & Oja, 2000; Beckmann et al., 2005) and Hidden Markov models (HMM, Rabiner, 1989; Vidaurre et al., 2016).

In non-deterministic approaches such as ICA and HMM, the degree to which different initialisations will lead to different estimations (i.e. different local minima) of the model parameters depends on elements such as the signal-to-noise ratio, training parameters, and amount of available data (Himberg et al., 2004). Many of the local minima may be equally good, or the inference may not

always perform well, for example getting stuck in suboptimal local minima. Either way, in order to assess how accurately the model estimates features of the data that cannot be observed directly (e.g. how much time a participant spends in a particular covert state during a period at rest), we can run the inference several times and, separately, perform permutation testing to assess the statistical features for each run. However, the goal of our analysis is not to identify features of a specific run (or specific experiment replications) but to approximate the 'true' underlying value that is (noisily) measured by the chosen approach, e.g. the HMM or ICA. To effectively estimate this underlying value, it is necessary to be able to combine the tests performed on multiple inference runs into a single global test.

Leaning on previous work (Winkler et al., 2016), this paper presents a simple approach to the problem of combining results from multiple runs using the principles of permutation testing, regardless of whether the replications are at the level of data acquisition, or model inference. This approach is useful in estimating effects that explain the underlying data that is the focus of the analysis. We demonstrate the validity of this method on the HMM, using simulations and data from the Human Connectome Project (Smith et al., 2013), where we test a measure of (resting state fMRI) dynamic functional connectivity over one hundred different HMM runs against a number of behavioural variables measured across hundreds of subjects.

Methods

Background

We refer to the noisy samples or parameter inference runs as *R replications*, to be distinguished from the *P observed variables* against which we aim to test. (Replications are not to be confused with *realisations*, which will use to refer to the multiple instances of the synthetic experimental scenario carried out below.) That is, we have one hypothesis per observed variable, and wish to combine the tests across multiple replications, with no particular interest in assessing each replication in isolation. For N subjects, let us denote replications as $\boldsymbol{X}$ (N by R), and observed variables as $\boldsymbol{Y}$ (N by P). For reference, we will consider each column of $\boldsymbol{X}$ (referred to as $\mathbf{x}_i$) as a noisy sample of certain unobservable variable of interest x.

For each column of $\boldsymbol{X}$ and each column of $\boldsymbol{Y}$ (referred to as $\mathbf{y}_i$), we can use permutation testing (Nichols and Holmes, 2002) to test the null hypothesis that $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ are independent (i.e. that there is no relationship between the estimated model and observed data). From this procedure we obtain a (1 by R) vector of p-values per observed variable, say $\boldsymbol{p}_j$. A naive approach could combine these R values with a simple statistic such as the mean or the median of $\boldsymbol{p}_j$ to assess the significance: if the mean p-value is small (e.g. below 0.01), this would suggest that there is a significant relationship between x and $\boldsymbol{y}_j$. In what follows, we will refer to this summarised p-value as p_{mean}. Alternatively to the mean, we can use the geometric mean, equivalent to exponentiating the average of the log p-values; this is related to Fisher's p-value combining method amplifies the

importance of values near zero. Denoting the individual p-values for a given observed variable of interest as p_i, we have

$$(1)\ p_{gmean} = \exp\left(\Sigma_i \log(p_i) / R\right).$$

Again, if p_{gmean} is below 0.01, we can state there is a significant relationship between the replications and the examined observed variable. Note that, although both p_{mean} or p_{gmean} are reasonable ways to combine tests, neither are rightful p-values, as they do not distribute uniformly in [0,1] under the null. These approaches, therefore, will work effectively as long as there is a consistent effect in all or most replications, but will fail if the models produce noisy estimates of the the data, or when only a subset of the models are able to capture any relationship with $\boldsymbol{y}_j$ at all.

Example case for a single pair of variables

Before coming to the description of the proposed approach, let us consider an example, where we wish to assess if there is a linear relationship between two variables, $\boldsymbol{a}$ and $\boldsymbol{b}$. The first one, $\boldsymbol{a}$, with values a_t, is Gaussian distributed (mean 0, standard deviation 1); the second one, $\boldsymbol{b}$, has elements

$$b_t = \kappa\, a_t + \varepsilon_t, \quad \text{for } t = 1 \ldots T,$$

where ε_t is Gaussian distributed (mean 0, standard deviation 1), and $\kappa > 0$ is randomly sampled from a uniform distribution between 0 and some pre-defined value c; here, c is manually chosen to define the expected strength of the relationship between $\boldsymbol{a}$ and $\boldsymbol{b}$. We sample from this distribution a number of times (or replications), each with a different value of κ that defines how strong the relation is between $\boldsymbol{a}$ and $\boldsymbol{b}$ for a given replication. We then run permutation testing on each sample data set. If c is higher than zero, then permutation testing analysis will be able to detect a significant relationship for at least some of the realisations (when κ is large); however, as long as c is not too large, it will not detect a significant relationship for some other realisations, i.e. those where κ is too small and the relationship between $\boldsymbol{a}$ and $\boldsymbol{b}$ is too weak to yield a significant p-value.

For the purpose of illustration we generated 1000 data sets using T=100, each with a different value of κ sampled from a uniform distribution, and performed permutation testing for each of them. We repeat this for three different values of c: 0.0, 0.1 and 0.2. **Figure 1** shows histograms of correlation coefficients between $\boldsymbol{a}$ and $\boldsymbol{b}$ across data sets (top), and histograms of p-values (bottom). If the empirical distribution of p-values is basically flat, as is the case when c=0.0, then there no evidence of a relationship between between $\boldsymbol{a}$ and $\boldsymbol{b}$. However, when *c=0.1* or *c=0.2*, then the distribution of p-values gets increasingly skewed toward zero. Therefore, if $\boldsymbol{a}$ and $\boldsymbol{b}$ were experimental replications of some pair of unobserved processes, we could intuitively say that there are signs of correlation between these processes in the c=0.1 and c=0.2 cases. However, neither p_{mean} or p_{gmean} is below 0.05 (they are all higher than 0.2 in all cases).

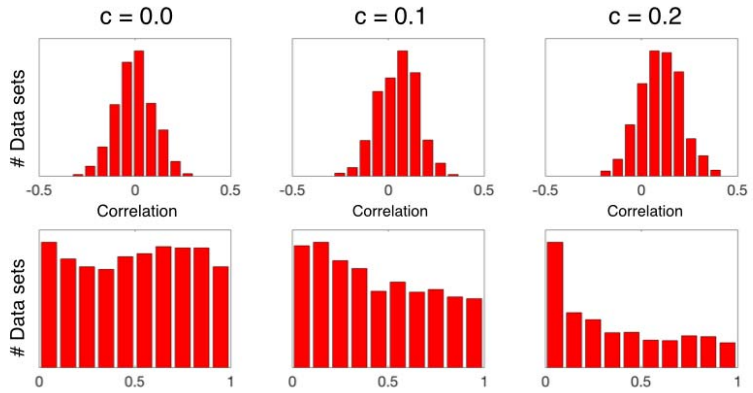

Figure 1: Simulated examples where we generated 1000 data sets, each consisting of two variables whose true correlation, c, is systematically varied. When c>0.0, the mean correlation across data sets is higher than zero (top), and the distribution of p-values is skewed toward 0.0 (bottom). The mean p-value (either in linear or logarithmic space) is higher than 0.05 in all cases.

Nested permutation testing

In the prior scenario, we are not interested in individual replications but on the relationship between the underlying variables (of which the replications are noisy observations). In this situation, it is clear that using the mean and median as a way to summarise the distribution of p-values can lead to Type II errors. In what follows, we propose a *nested* permutation testing procedure to overcome this limitation. This procedure follows from the principles presented by Winkler et al. (2016).

In the case when there is only one observed variable (P=1), referred to as $\boldsymbol{y}$, we propose the following algorithm:

Algorithm 1

I. At the first level, we perform permutation testing under the null hypothesis of independence between each replication $\boldsymbol{x}_i$ and $\boldsymbol{y}$ to obtain an empirical distribution of p-values, represented by the (R by 1) vector of p-values $\boldsymbol{p}^0$ (e.g. Figure 1-middle). We can implement this by randomly permuting $\boldsymbol{x}_i$ a number of times, and comparing the surrogate correlations with that of the unpermuted data.

II. We summarise $\boldsymbol{p}^0$ using the geometric mean, which, using Equation (1), yields p_{gmean}. This corresponds to the first-level permutation testing.

III. For k = 1 ... K (where K is some large number), we apply some valid permutation on $\boldsymbol{y}$ to create a null hypothesis surrogate $\boldsymbol{y}^k$ (under the same null hypothesis of independence than step I). For each k, then, we repeat steps I and II, now using $\boldsymbol{y}^k$ instead of $\boldsymbol{y}$, such that we get vectors of p-values $\boldsymbol{p}^k$ and combined p-values p^k_{gmean}.

IV. At the second level, we obtain a final p-value as

$$(2)\ p_{nested} = (\#_k\{p_{gmean} \geq p^k_{gmean}\} + 1) / (K+1).$$

We can easily extend this procedure for the $P > 1$ case (when there are more than one observed variable of interest) by using Equation (1) on each observed variable $\boldsymbol{y}_j$ separately while using the same permutation scheme for all observed variables, such that the dependence between the tests is implicitly accounted for (Winkler et al., 2016). This will yield a final p-value per observed variable, say $p_{nested,j}$. We can obtain a summary, family-wise error corrected p-value (Nichols and Hayasaka, 2003) for each variable of interest j by computing

$$(3)\ p^{\mathrm{FWE}}_{nested,j} = (\#_k\{p_{gmean,j} \geq \min_j(p^k_{gmean,j})\} + 1) / (K+1),$$

where $p^k_{gmean,j}$ is the null surrogate p-value obtained with Equation (1) for the j^{th} variable of interest and k^{th} realisation. Alternatively, we can use false-discovery rate (FDR; Benjamini and Hochberg, 1995; Nichols and Hayasaka, 2003) on the uncorrected p-values $p_{nested,j}$ to obtain FDR-corrected p-values $p^{FDR}_{nested,j}$.

Therefore, if there is any relationship between x and $\boldsymbol{y}_j$, the empirical distribution of p-values represented by $\boldsymbol{p}^0$ will be, at least to some extent, skewed to the left (as in **Figure 1**, bottom row, centre and right columns) even if the mean, or the geometric mean, are not under the area of statistical significance; in contrast, the empirical null distribution for the permutations will be largely flat (as in **Figure 1**, bottom left).

If computational cost is a concern, an alternative is to use some form of (fast) parametric testing at the first level instead of permutation testing. As shown empirically below, the use of second-level permutation reduces the consequence of violations of the assumption of normality.

In summary, this procedure draws statistical power from both working in logarithmic space (i.e. promoting the importance of p-values closer to zero, Winkler et al., 2016), while simultaneously relaxing the alternative hypothesis from a highly conservative assumption – that "*most* of the replications bear a relationship with the corresponding observed variable" to a less conservative assumption that "*at least* some of the replications bear a relationship with the corresponding observed variable". In the above example, for instance, nested permutation testing produced a p-value higher than 0.5 when c=0.0, and p-values lower than 0.001 for both the c=0.1 and c=0.2 cases, exhibiting both robustness and sensitivity.

Simulations

To illustrate this method, we simulated synthetic data sets emulating a scenario in which we are interested in testing whether functional connectivity (FC) between a pair of brain regions holds a relation to certain behavioural trait in a set of N subjects. In this situation, we have the following variables:

- A subject-specific FC coefficient β, which we cannot observe directly.
- A behavioural variable, hypothesised to be related to FC and encoded by a (N by 1) vector $\mathbf{y}$, that can be observed directly.
- Some neural process modulated by β, denoted as $\boldsymbol{S}$, which we cannot observe directly. We can consider $\boldsymbol{S}$ to be some archetypical, noiseless brain activity controlled by β.
- The observed (e.g. neuroimaging) data sets $\boldsymbol{D}$, which are noisy measurements of $\boldsymbol{S}$ and have dimension (T by 2). This measurement can be repeated up to R times per subject.
- An (N by R) matrix $\boldsymbol{X}$, such that X_{ni} contains the estimated FC value for the n^{th} subject and i^{th} experimental replication (i.e. the correlation coefficient between the channels of the corresponding measured data $\boldsymbol{D}$).

In this context, the noise in the observations (or replications) stems from the imperfect measurement of $\boldsymbol{S}$, which we can measure multiple times (R). Therefore, there is a relation between FC (β, which we cannot observe but we can estimate) and behaviour ($\boldsymbol{y}$), but this relationship is noisy and weak for some replications. The objective of this simulation is then to assess whether the proposed approach can uncover such relationship. In detail, we generate data from this setting as follows:

We have N=200 subjects. We uniformly sample a value β_n between -0.2 and +0.2 for each subject n. For each subject, also, we sample two vectors with 10000 values each: the first, $\boldsymbol{s}_1$, is Gaussian distributed (mean 0, standard deviation 1), whereas the second is set as

$$\boldsymbol{s}_2 = \beta_n \boldsymbol{s}_1 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{\varepsilon}$ is also Gaussian-distributed. The vectors $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ constitute the unobserved neural process $\boldsymbol{S}$. The correlation between $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ can be analytically computed from β_n as

$$c_n = \beta_n / (\beta_n^2 + 1)^{1/2}$$

We set the value of the observed behavioural variable for each subject to be

$$y_n = c_n + 0.5\, \eta_n,$$

where $\boldsymbol{\eta}$ is Gaussian distributed (mean 0, standard deviation 1). Now, in order to sample the observed data sets $\boldsymbol{D}$ for each subject, we randomly sample 100 time points from $\boldsymbol{S}$ (whose columns are $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$) and add some Gaussian noise with mean 0 and standard deviation σ. We do this R times per subject, obtaining one (100 by 2) noisy data set $\boldsymbol{D} = [\boldsymbol{d}_1, \boldsymbol{d}_2]$ each time. We then set the observed replication values to

$$X_{ni} = \text{corr}(\boldsymbol{d}_1, \boldsymbol{d}_2)$$

With both $\boldsymbol{X}$ and $\boldsymbol{y}$ in hand, we run the proposed nested permutation testing algorithm on the noisily estimated FC matrix $\boldsymbol{X}$ and the behavioural variable $\boldsymbol{y}$: that is, as described in Methods, in the first level (I) we permute the rows of $\boldsymbol{X}$, and in the second-level (III) we permute the elements of $\boldsymbol{y}$. By controlling σ (which defines how noisy are individual time series samples $\boldsymbol{d}_1$ and $\boldsymbol{d}_2$), we can make the problem more or less difficult. We use a range of 30 values for σ between 0.25 and 1.5, and repeat data generation and testing 100 times per value of σ. For each repetition of the experiment, standard permutation testing results on R=100 p-values (one per replication). Note that, since P=1, there is no need to control for familywise error rate across observed variables (Equation (3)).

On top, **Figure 2** shows p_{mean} / p_{gmean} / p_{nested} (respectively from left to right) averaged across the 100 realisations of the experiment as a function of σ, together with 95% confidence intervals (minus/plus twice the standard error). Thanks to the effect of the logarithm, the p_{gmean} values are lower than p_{mean} values, but neither of them ever reach significance provided the weak and volatile relationship between $\boldsymbol{X}$ and $\boldsymbol{y}$. The individual, per replication p-values (shown underneath for one example repetition, per value of σ) illustrate this point: although there are some significant p-values, the average is condemned to fail due to the frequent bad p-values associated to some too noisy replications. However, most of the p-values from the nested permutation approach turned out to be significant despite the difficulty of the problem, with the average of p_{nested} across realisations leaving the zone of significance only for the highest values of σ (i.e. for the hardest instantiations of the problem).

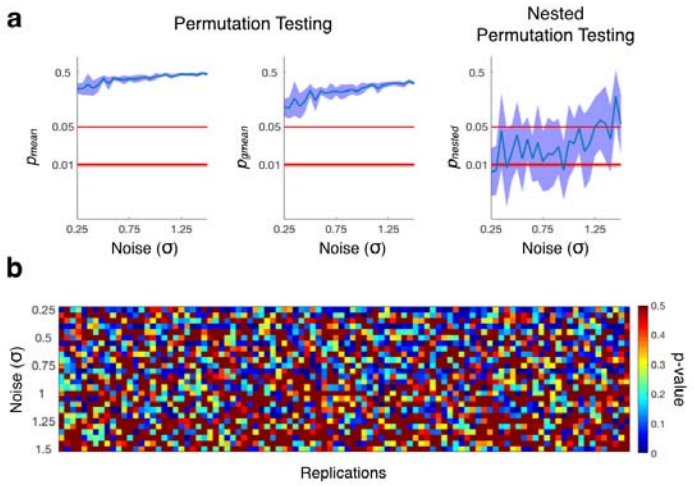

Figure 2. Results from the simulated data, where there is a relationship between the tested variables: FC and behaviour. (**a**) p-values obtained from combining tests by using the mean (p_{mean} and p_{gmean}), and p-values from the proposed nested permutation testing approach (p_{nested}), as a function of σ, which controls the noise in the replications (i.e. higher values of σ produce more difficult instantiations of the problem); intervals of confidence are computed across realisations of the experiment. (**b**) p-values before test combination for a given repetition (per value of σ).

Next, we repeat the same analysis but forcing a fixed value of β_n for all subjects (in particular, we set $\beta_n = 0$). In this case, there is relationship between behaviour and FC. Figure 3 shows that nested permutation testing, as well as the other methods, are robust and do not yield Type I errors in this scenario.

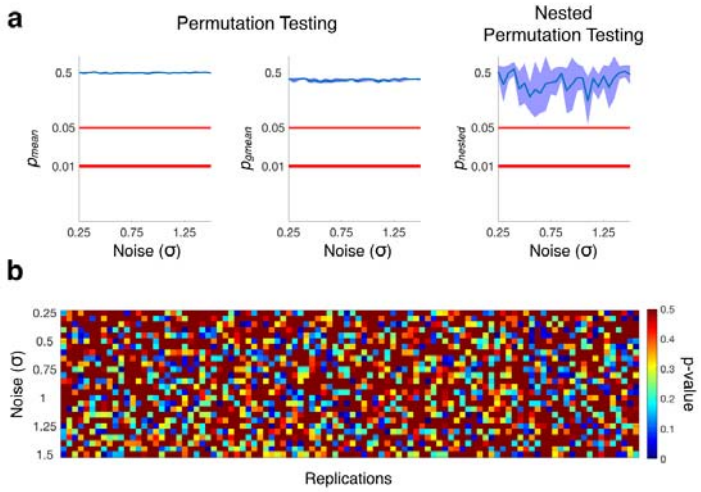

Figure 3. Results from the simulated data, where there is *not* a relationship between FC and behaviour. The description of the panels is equivalent to Figure 2. In this case, however, no relation was found between FC and behaviour, i.e. there are no Type I errors.

Dynamic functional connectivity in real data

Having demonstrated the utility of the nested permutation approach using synthetic data, we next evaluate it using real data by applying the Hidden Markov model (HMM) to resting state fMRI data from the Human Connectome Project (HCP). The HMM assumes that the data can be described using a finite number of states. Each state is represented using a probability distribution, which in this case is chosen to be a Gaussian distribution (Vidaurre et al, 2017a); that is, each state is described by a characteristic pattern of BOLD activation and certain functional connectivity profile (we use the same configuration as in Vidaurre et al (2017a), to which we refer for further details). As the HMM is applied at the group level, the estimated states are shared across subjects; however, the state time courses that indicate the moments in time when each state is active are unique to a given individual. For the purposes of this analyses we set the HMM to have 12 states. Using the inferred state time courses, the amount of *state-switching* for each subject is calculated (Cabral et al, 2017), which corresponds to a metric of how frequently subjects transition between different brain states (more specifically, given that the state time courses are probabilistic assignments, we compute the mean derivative of the state time courses for each subject). We use state switching as a summary metric of dynamic functional connectivity (DFC).

In order to infer the HMM at reasonable cost in spite of the large amount of data (820 subjects by 4 sessions by 15min, TR=0.75s), we use a stochastic learning procedure (Vidaurre et al, 2017b), which involves performing noisy, yet economical, updates during the inference. Since stochastic inference brings an additional layer of randomness into the HMM estimation but is not costly to run, we repeated the HMM inference 100 times and computed state-switching for each run. In this context, each HMM estimation constitutes a replication.

Furthermore, each subject has a number of behavioural measures, including psychological and sociological factors and several health-related markers. We used a total of 228 behavioural variables, after discarding those with more than 25% of missing values, in order to test against DFC as measure by state-switching. We included age, sex, motion and body-mass-index (the latter two usually considered as confounds). We also discarded those subjects without family information (important for creating the permutations; Winkler et al, 2015), and those with a missing value in any of the behavioural variables.

Although stochastic inference adds additional randomness to the estimation, the HMM has have previously been reported to perform very robustly in this data set (Vidaurre et al. 2017a), possibly as a consequence of the large number of subjects (N=820), the length of the scanning sessions, and the general high quality of the data. For this reason, the different HMM runs are quite consistent, which in turn means that the tests produce relatively similar results across replications (as shown below). To illustrate the effect of greater noise, we created a second set of replications where we permuted the state-switching measure between subjects randomly for half of the HMM runs (that is, half of the HMM runs, or replications, are potentially related to behaviour whereas the other half are noise, and all of them are included in the analysis). We refer to this as the *perturbed* data, as opposed to the *original* data where the HMM estimations are left intact.

Figure 4 compares the results of applying the nested permutation testing approach with the mean p-value either in logarithmic or linear space. We used 10000 permutations in the outer loop, and 1000 permutation in the nested loop to obtain each p^k_{gmean} p-value. **Figure 4a** shows the mean p-value (averaged across replications) reflecting the subject-wise correlation of state-switching (as measured by the HMM) with the different behavioural variables, with the behavioural variables being ordered from more to less significant; dots represent individual p-values for some randomly chosen replications. On the left, the p-values obtained from standard permutation testing on the original HMM runs are quite consistent across replications; on the right, for the perturbed set of HMM runs, given that half were randomly ordered over subjects, the mean p-value is hardly significant for any of the variables.

We next compute the cumulative distribution of p-values for the 62 variables that reached significance using all methods, used as a summary of performance. **Figure 4b** demonstrates the advantage of using nested permutation testing. On the left, where all the HMM runs were used normally, this difference is subtle; on the right, however, the difference is substantial. The difference between p_{mean}

and p_{gmean} conveys the benefits of working on logarithm space, whereas the difference between p_{gmean} and p_{nested} reflects the increased sensitivity brought about by testing the "right" hypothesis ("most replications" vs "at least some replications" relate to behaviour). **Figure 4c** shows, for each of the methods, the (combined across replications) p-values for the original data versus the perturbed data, reflecting that only the nested permutation testing approach is robust to having corrupted replications (i.e. the p-values are almost identical between the original and the perturbed data set).

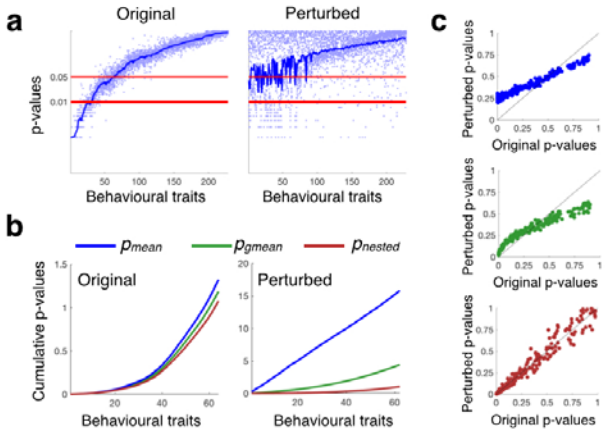

Figure 4. Analysis of the relation between behaviour and DFC (state switching) as measured by the HMM, where replications correspond to HMM runs. (**a**) Mean p-values (averaged over replications, with dots representing p-values for some individual replications), reflecting the subject-wise correlation of DFC with the different behavioural variables. On the X-axis, behavioural variables are ordered from more to less correlated. On the left, this is shown for the original data set; on the right, this is shown for the perturbed data set (a noisier version of the original data set). (**b**) Nested permutation testing outperforms p_{mean} and p_{gmean}, as reflected when we examine the cumulative distribution of p-values. (**c**) The p-values are robust to perturbation only for the nested approach, where correlation between perturbed and original p-values is close to 1.0.

Figure 5 presents the behavioural variables for which we found significance using the nested permutation testing procedure. Interestingly, although motion is a significant predictor it does not explain the greatest variance in this analysis, suggesting that DFC on resting state fMRI, as estimated by HMM, can be meaningfully related to behaviour beyond the influence of motion. Due to the relatively large number of observed variables, FWE correction is fairly conservative and few observed variables turn out to be significant (that is, in Equation (3), the minimum of the surrogate p-values across observed variables can be small if there are many observed variables to choose from). In contrast, FDR, being a less conservative approach that FEW correction (Nichols & Hayasaka, 2003), preserves statistical significance for up to 26 variables. Importantly, if we randomly corrupt the entire data set (instead of half of the subjects as in the perturbed data set), all methods, including nested permutation testing, are able to avoid Type I errors by marking all behavioural variables as not significantly related to DFC (not shown).

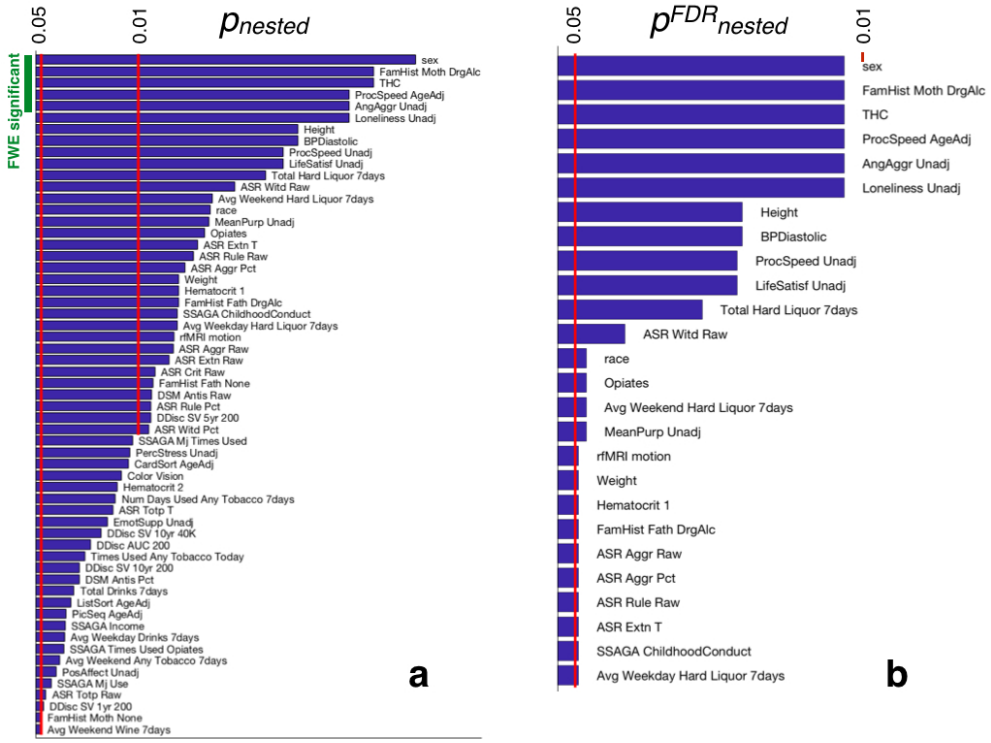

Figure 5. For the observed variables considered to be significant (out of 228), (**a**) p-values using the nested permutation testing approach (p_{nested}), with FWE significance indicated on the top left; and (**b**) FDR-corrected p-values ($p^{\mathrm{FDR}}{}_{nested}$).

Finally, to assess the impact of using parametric testing instead of permutation testing at the first level, we repeated the analysis using t-testing in step I of **Algorithm 1**. The results, shown in **Figure 6**, are comparable, suggesting that parametric testing can be reasonably used as a replacement of permutation testing at the first level when computations are too costly.

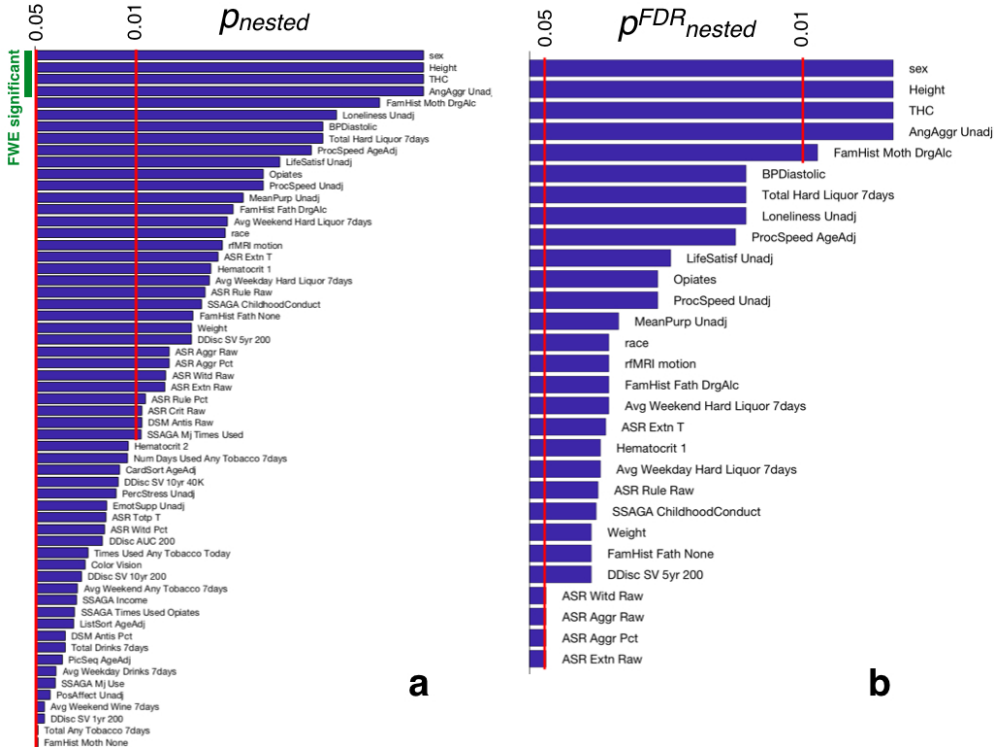

Figure 6. Similarly as shown in Figure 4 but using parametric instead of permutation testing in the first level (step I in the nested permutation testing procedure), (**a**) p-values using the nested permutation testing approach (p_{nested}), with FWE significance indicated on the top left; and (**b**) FDR-corrected p-values ($p^{FDR}{}_{nested}$).

Discussion

Based on previous work (Winkler et al., 2016), we propose in this paper an approach for testing for the relationship between a set of observed variables and an unobserved variable for which we have a number of noisy measurements. In this case, we are not interested in estimating the relationships as described by a particular measurement or replication, but instead would like to understand the relationship of the *true* unobserved variable with the observed variables. We took as a concrete example the relationship between covert patterns of intrinsic brain connectivity, as they occur at rest, and patterns of cognitive and demographic variables measured outside of the scanner, using data from the HCP data. In this example, the patterns of intrinsic connectivity reflect the noisy unobserved data, whereas the measures of cognitive functioning reflect measureable observed variables. We demonstrated, using both synthetic and real data, that nested permutation testing is able to identify real relationships in a manner that is less likely to make a Type II error than would be possible using other alternatives.

Although we focused on univariate observed variables and replications, the proposed method can straightforwardly be extended in a number of ways. First, although we focused on linear relationships between variables, it can easily be extended to multivariate statistics, such as multivariate linear regression, or

canonical correlation analysis. This will be important because it will allow the exploration of situations in which the mapping between cognitive function and the data is not univariate in nature. It can also be extended to situations when we have replications on both sides of the correlation, such as when both the observed and non-observed behaviours are measured on multiple occasions. In this case, each pair of replications could be tested individually (for each element of the corresponding Cartesian product), and we would proceed in a similar fashion.

Moving forward, our use of nested permutation testing is likely to be particularly important in the domain of neuroscience given recent shifts towards the use of intrinsic connectivity at rest as a method of evaluating structural features of cognition. Intrinsic connectivity, as measured at rest, is a powerful tool for exploring the structure of neural organisation since it is able to reveal similar patterns of neural organisation as emerge during tasks (Smith et al., 2009). In addition, the simple non-invasive nature of the use of resting state as a method for assessing neural function means that it can be applied to multiple different populations, even those for whom task based measures of neural function, or psychological measurements may be problematic (such as children or populations with cognitive problems). Measuring neural organisation at rest is also easy to implement across centres making it amenable to the creation of large multicentre data sets, a shift that is likely to be increasingly important as neuroscience faces up to the challenges of reproducible science.

Despite the promise that assessing neural function at rest holds, many of the same features that make it an appealing tool for the cognitive neuroscience community are also at the heart of its significant limitations. For example, the power that is gained by the unobtrusive nature of the measure of neural function at rest also leads to concerns regarding what the measures actually reflect: it is unclear which aspects of the neural signal reflect the intrinsic organisation of neural function, which reflect artefacts that emerge from physiological noise or motion (Power et al., 2013), and which reflect the patterns of ongoing experience that frequently emerge when individuals are not occupied by a demanding external task (Gorgolewski et al., 2014, Vatansever et al., 2017). In this context, because the underlying ground truth is unknown, nested permutation will help the researcher to identify which aspects of a given neural pattern are expressed in a robust way in relation to neurocognitive function.

Similarly, given its lack of constraints, it remains unclear whether the resting-state should be treated as a single continuous state, as it is frequently assumed when simple measures of functional connectivity are used, or whether it should instead be treated as a sequence of dynamic states (Chang and Glover, 2010; Vidaurre et al., 2017a). Although dynamic approaches to understanding functional connectivity space are growing in popularity, different approaches have specific limitations. For example, sliding window approaches depend upon an apriori selection of the window length, which limits the granularity of neurocognitive states that can be identified. While approaches such as HMM circumvent this problem by allowing the data to determine the temporal duration of the underlying states, these analyses are inherently probabilistic and

parameter inference can introduce noise into the analysis. In this context, nested permutation testing allows dynamic approaches to cognition to be compared to observed data in a systematic manner. This could help paving the way to formally evaluate how different descriptions of the underlying dynamics at rest best predict variables with well-described links to cognitive function. This way, nested permutation testing can become a useful tool in resolving the state-trait dichotomy that currently hinders the development of the science of how neural function evolves at rest.

Acknowledgements: The Wellcome Centre for Integrative Neuroimaging is supported by core funding from the Wellcome Trust (203139/Z/16/Z). DV is supported by a Wellcome Trust Strategic Award (098369/Z/12/Z). MWW is supported by the Wellcome Trust (106183/Z/14/Z) and the MRC UK MEG Partnership Grant (MR/K005464/1). JS was supported by ERC (WANDERINGMINDS - 646927)

Bibliography

A. Hyvarinen & E. Oja (2000). Independent component analysis: algorithms and applications. *Neural Networks* **13**: 411-430

C.F. Beckmann, M. DeLuca, J.T. Devlin & S.M. Smith (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transitions of the Royal Society B* **360**: 1001-1013

L.R. Rabiner (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**: 257-286.

D. Vidaurre, A.J. Quinn, A.P. Baker, D. Dupret, A. Tejero-Cantero & M.W. Woolrich (2016). Spectrally resolved fast transient brain states in electrophysiological data. *NeuroImage* **126**: 81-95.

J. Himberg, A. Hyvärinen & F. Exposito (2004). Validating the independent components of neuroimaging time series via clustering and visualization. *NeuroImage* **22**, 1241-1222

M.J. Wainwright & M.I. Jordan (2008). Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning* **1**, 1-305.

S.M. Smith, C.F.Beckmann, J. Andersson, E.J. Auerbach, J. Bijsterbosch, G, Douaud, E. Duff, D.A. Feinberg, L. Griffanti, M.P. Harms, M. Kelly, T. Laumann K.L. Miller, S. Moeller, S. Petersen, J. Power, G. Salimi-Khorshidi, A.Z. Snyder, A.T. Vu, M.W. Woolrich, J. Xu, E. Yacoub, K. Ugurbil, D.C. Van Essen & M.F. Glasser. (2013) Resting-state fMRI in the Human Connectome Project. *NeuroImage* **80**: 144-168.

T.E. Nichols & A.P. Holmes (2002). Nonparametric Permutation Tests For Functional Neuroimaging : A Primer with Examples, *Human Brain Mapping*, **15**: 1–25.

Y. Benjamini & Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach for multiple testing. *Journal of the Royal Statistical Society B*, **57**: 289-300.

T.E. Nichols & S. Hayasaka (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, **12**: 419–446.

D. Vidaurre, S.M. Smith & M.W. Woolrich (2017a). Brain networks are hierarchically organised in time. *Proceedings of the National Academy of Sciences of the USA*, In press.

D. Vidaurre, R. Abeysuriya, R. Becker, A.J. Quinn, F. Alfaro-Almagro, S.M. Smith & M.W. Woolrich (2017b). Discovering dynamic brain networks from Big Data in rest and task. *NeuroImage*. In press.

J. Cabral, D. Vidaurre, P. Marques, R. Magalhães, P.S. Moreira, J.M. Soares, G. Deco, N. Sousa & M.L. Kringelbach (2017). Cognitive performance in healthy older adults relates to spontaneous switching between states of functional connectivity. *Scientific Reports* **7**: 5135.

A. Winkler, M.A. Webster, D. Vidaurre, T.E. Nichols & S.M. Smith (2015). Multi-level block permutation. *NeuroImage* **123**: 253-268.

A. Winkler, M.A. Webster, J.C. Brooks, I. Tracey, S.M. Smith & T.E. Nichols (2016). Non-parametric combination and related permutation tests for Neuroimaging. *Human Brain Mapping* **37**, 1486-1511.

S.M. Smith, P.T. Fox, K.L. Miller, D.C. Glahn, P.M. Fox, C.E. Mackay, N. Filippini, K.E. Watkins, R. Toro, A.R. Laird & C.F. Beckmann (2009). Correspondence of the brain's functional architecture during activation and rest. *Proceedings of the National Academy of Sciences of the USA* **106**, 13040-13045.

K.J. Gorgolewski , D. Lurie, S. Urchs, J.A. Kipping, R.C. Craddock, M.P. Milham, D.S. Margulies & J. Smallwood (2014). A Correspondence between Individual Differences in the Brain's Intrinsic Functional Architecture and the Content and Form of Self-Generated Thoughts. *PLoS One* **9**, e97176.

D. Vatansever, D. Bzdok, H. Wang, G. Mollo, M. Sormaz, C. Murphy, T. Karapanagiotidis, J. Smallwood & E. Jefferies (2017). Varieties of semantic cognition revealed through simultaneous decomposition of intrinsic brain connectivity and behaviour. *NeuroImage* **158**, 1-11.

C. Chang & G.H. Clover (2010). Time–frequency dynamics of resting-state brain connectivity measured with fMRI. *NeuroImage* **50**, 81-98.